Mathematical methods for marine energy extraction

Sebastián Riffo

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Université Paris-Dauphine

Méthodes mathématiques pour l'extraction d'énergie marine

Soutenue par
Sebastián REYES RIFFO
Le 29 novembre 2019

École doctorale n°543
École Doctorale de Dauphine

Spécialité
Sciences

Composition du jury :

M. Michel BERGMANN INRIA Bordeaux Sud-Ouest	*Rapporteur*
Mme. Mireille BOSSY INRIA Sophia Antipolis	*Présidente du jury*
M. Olivier GLASS Université Paris-Dauphine	*Examinateur*
M. Stefan GÜTTEL University of Manchester	*Rapporteur*
M. Philippe MOIREAU INRIA Saclay	*Examinateur*
M. Jacques SAINTE-MARIE INRIA Paris	*Examinateur*
M. Julien SALOMON INRIA Paris	*Directeur de thèse*

Dauphine | PSL
UNIVERSITÉ PARIS

"La Revolución no pasa por la universidad, y esto hay que entenderlo;
la Revolución pasa por las grandes masas;
la Revolución la hacen los pueblos;
la Revolución la hacen, esencialmente, los trabajadores."

Salvador Allende, Presidente de Chile (1970-73).

Acknowledgments

Working on this thesis was equivalent to living in a roller coaster during almost 4 years. Exciting times, I must admit, in which I visited amazing places as Svalbard or Hong Kong while enjoying my work, and I also had a life in Paris I've never dreamed of. Even so, in an almost unnoticeable way, I began to forget why I chose to be a mathematician, since that willingness to learn slowly transformed into a distaste for what I was doing. At some point, near to what was supposed to be the end, I was only worried I wouldn't finish this manuscript in time, but then I simply faced reality and said : if Brexit was postponed several times, why can't I do the same ? Still, sometimes it seemed closer than my long-awaited thesis defense, even Notre-Dame got burned down in between, and I was also afraid that GRRM, a professional procrastinator, could publish his last book at any moment. By the end this PhD was a mix between uneasiness, impostor syndrome, total relief and more than a bit of hair turning white, but still, we made it.

You may be wondering why I took the liberty of writing all this. It's just that we inevitable soften our memories, and among many other things, writing is our small rebellion against time. In any case, it's time to come back to the usual purpose of these lines. I would like to thank all of whom were part of this process, because I couldn't have finished this thesis alone. Sure enough, an enormous cliché.

⁂

Since protocol always goes first, let me start with the members of the jury, possibly the only readers of this manuscript. I'm deeply indebted to Julien Salomon, my advisor, first of all for trusting me from the very beggining (even to the point of sending my *dossier* to the last PhD scholarship he found, which I got !), and then for guiding me through this whole process. I truly appreciate his patience and support at those critical times when I felt my research was falling apart, as in *Le Radeau de la Méduse*. I would like to thank both of my reviewers, Stefan Güttel and Michel Bergmann, for their constructive criticism and suggestions that helped me to improve this work. I'm also grateful to Mireille Bossy, Olivier Glass, Philippe Moireau and Jacques Sainte-Marie for agreeing to be part of my committee.

I'm extremely grateful to Felix Kwok, with whom I was lucky enough to collaborate in one of the topics covered in this thesis. Both research visits to Hong Kong were quite instructive, not only at the mathematical level, since I learned what it's really enjoy researching and make it compatible with everyday life. I also had the great pleasure of working with Pierre-Henri Cocquet and Jérémy Ledoux.

I would also like to extend my gratitude to all members of CEREMADE. Even if the architecture of the university building wasn't appealing, as military complexes often are, I really enjoyed my stay there. Many thanks to my office mates Nadia, Fang, Long and the aircon machine, our invaluable summer companion. Working together was like having a support group, and I definitely learned a lot from our interesting conversations about each other's respective cultures ! Thanks also to Jorge and the rest of the ephemeral *pisco sour* gang. Speaking of offices, being four floors below the rest of the lab was an excellent way to have a healthier lifestyle, based on climbing stairs, drinking coffee afterwards and hopefully, turned it into theorems; which also reminds me it wasn't until last year that I realized coffee was free for us. Then I guess it's not a surprise to anyone that I wasn't aware of the usual –and most of the time, nightmarish– paperwork, so I gratefully acknowledge the assistance of all administrative people, especially Cesar Faivre and Isabelle Bellier, Linda Mammoudi from the École doctorale, as well as Vincent Rivoirard and Jacques Féjoz on the directive side.

⁂

Finalmente, en la vorágine de ideas, fórmulas y técnicismos propios de este doctorado, es fácil creer que sus resultados son un fruto más de la "meritocracia" y omitir por completo el rol que jugaron mi familia y amigos en toda esta historia. Me gustaría agradecerles, aunque sea brevemente, en las próximas líneas.

A Milena, por darme esa claridad tan propia de ella que siempre termina poniéndome en perspectiva, convencerme de aceptar una beca inesperada y ser parte fundamental de estos años intensos y viajeros. Que nuestro actual paso por Francia y el futuro nos deparen más cosas buenas! A mi mamá y hermanos, por su generosidad al apoyar incondicionalmente cada una de mis decisiones, muchas veces precipitadas, confiando que estaba en lo correcto y aún sin entender porqué decidí dedicarme a esto.

Gracias a los amigos matemáticos que decidieron repartirse por el mundo y de paso compartir un nuevo capítulo de esta estafa piramidal: Mónica y sus precisos consejos; Sandra y Jean-marc, por adoptarme cual gato en su hogar; Johan, Dani, señora Nicole y Abelino; Pedro y Pita, por las escapadas al *château*; y la capitana Anne, siempre más visionaria que todos nosotros. También a quienes de, a pesar de lo poco que nos vemos, nuestra amistad sigue intacta: Juan Carlos, Felipe, Karla, Gabriel, Pepo, por mencionar sólo a algunos. Y a los amigos con los cuales he compartido la experiencia de vivir acá, especialmente Víctor y Nico, Carlos y Garrido tiempo después. Las anécdotas del programa capital humano avanzado en París son interminables.

Contents

List of Figures

List of Tables

Résumé de la Thèse

Introduction Générale

La présente thèse vise à contribuer à l'élaboration d'un cadre théorique pour trois problèmes dans le contexte des énergies marines renouvelables, à savoir la parallélisation en temps de l'assimilation de données, l'optimisation d'une bathymétrie et l'analyse mathématique de la méthode de l'élément de pale (BEM). Puisque leur résolution repose en grande partie sur des connaissances empiriques, nous croyons que l'adoption d'un point de vue mathématique mène à une meilleure compréhension des différentes situations, ce qui constitue une occasion d'encourager la collaboration interdisciplinaire entre les mathématiques et les sciences appliquées comme la géophysique et l'ingénierie. Dans ce qui suit, nous présentons nos principales contributions sur chaque sujet.

Parallélisation en temps de l'assimilation de données

Les hypothèses qui sous-tendent un modèle mathématique déterminent non seulement leur plage d'application, mais elles induisent également un écart inévitable entre les prévisions et la réalité. Afin de réduire cette différence, nous pouvons incorporer des données réelles au lieu de sacrifier la simplicité du modèle, en suivant une procédure d'assimilation de données (AD). Plusieurs aspects doivent être pris en compte lors de l'application de ces techniques à e.g. des problèmes de météorologie ou d'océanographie, mais nous rappelons ici qu'en raison du nombre de variables d'état et du grand quantité d'observations requises, leur résolution numérique est coûteuse en temps de calcul. Trémolet et Le Dimet [88] ont été parmi les premiers à aborder la parallélisation des problèmes d'assimilation de données variationelle (qui sont basés sur la théorie du contrôle optimal et utilisent les informations collectées en un temps donné) en utilisant une approche de décomposition de domaine (DD). Depuis lors, le couplage entre ces deux procédures a été largement étudié.

D'une façon générale, les méthodes de DD consistent à décomposer la dimension spatiale en sous-domaines, avec un possible chevauchement, puis à résoudre de manière synchrone un problème local sur chacun d'eux. Cette stratégie de division et de conquête semble contre-intuitive lorsqu'il s'agit de gérer la direction temporelle, en raison de sa nature séquentielle inhérente, ce qui explique pourquoi la parallélisation en temps n'est pas couramment appliquée aux problèmes d'AD.

Les algorithmes de parallélisation en temps peuvent être très utiles lorsqu'il s'agit de intervalles de temps plus longs, comme c'est le cas des méthodes d'AD séquentielles, où les informations peuvent arriver sans interruption. Une question naturelle se pose alors : pouvons-nous combiner les deux procédures ? Nous commençons à répondre à cette question au Chapitre 2, en étudiant l'observateur de Luenberger et son couplage avec l'algorithme Pararéel.

Optimisation d'une bathymétrie

Bien que la bathymétrie puisse être mal connue dans de nombreuses situations, les modèles de propagation des ondes dépendent fortement de ce paramètre pour capturer le comportement de l'écoulement, ce qui souligne l'importance d'étudier les problèmes inverses concernant sa reconstruction à partir des données observées en surface libre.

Ces types de problèmes sont habituellement résolus en discrétisant simplement les équations qui les régissent ou à l'aide de méthodes d'AD séquentielles. Une alternative consiste à considérer la bathymétrie comme variable de contrôle d'un problème d'optimisation sous contrainte d'EDP, une approche utilisée en ingénierie côtière en raison des contraintes mécaniques associées aux structures des bâtiments et à leur interaction avec les vagues de mer. Cependant, sa résolution repose principalement sur des analyses de sensibilité, des méthodes numériques ou des simplifications du modèle qui conduisent à des solutions explicites et ensuite, les questions concernant un espace de contrôle approprié, la continuité de la fonction contrôle-état, la régularité des solutions et le caractère bien posé du problème ne sont en général pas abordées. Récemment, Dalphin et Barros [25] ont réalisé cette analyse théorique pour modéliser un générateur de vagues.

Au Chapitre 3 nous essayons de répondre aux questions ci-dessus lorsque nous envisageons une reformulation de l'équation de Helmholtz pour la modélisation de la propagation des ondes et une fonctionelle générale de coût qui peut être identifiée, par exemple, avec le décalage entre la solution d'onde prévue et celle observée.

Analyse mathématique de la méthode de l'élément de pale

Hydrotube Énergie, entreprise dédiée à la conception nautique et aux systèmes électroniques embarqués, a testé en 2015 le prototype d'une turbine hydraulique flottante sur la Garonne, à Bordeaux. Mais l'évolution vers une production à l'échelle industrielle nécessite le développement d'un logiciel numérique pour simuler un dispositif optimisé, qui est le but de son partenariat avec l'équipe ANGE (INRIA), un groupe de recherche en modélisation, analyse et simulation des écoulements géophysiques.

Parmi les problèmes à résoudre, nous étudions l'optimisation de l'efficacité de la turbine via la *méthode de l'élément de pale*. Présenté par Glauert [44], cette méthode est largement utilisée pour déterminer l'efficacité et donc les paramètres de conception d'une pale, en fonction de ses caractéristiques géométriques et mécaniques et du courant auquel elle est exposée. Elle résulte de la combinaison de deux méthodes indépendantes qui traitent le système fluide/turbine d'un point de vue macroscopique et local, ce qui donne un ensemble d'équations qui sont résolues en appliquant un algorithme itératif. Dans certains cas, un facteur de correction est nécessaire pour assurer l'existence des solutions du modèle.

Bien qu'ancienne, cette méthode est encore utilisée en raison de sa relative simplicité par rapport à la complexité du phénomène hydrodynamique développé dans le système fluide/turbine. Néanmoins, l'existence de solutions et la convergence de l'algorithme n'ont jamais été analysées d'un point de vue mathématique, ce qui est le but du Chapitre 4.

Contributions de cette thèse

Ce travail est divisé en trois chapitres indépendants portant sur les sujets susmentionnés, dans lequel j'ai eu le plaisir de collaborer avec Julien Salomon (INRIA Paris), Felix Kwok (Hong Kong Baptist University), Pierre-Henri Cocquet (Université de Pau et des Pays de l'Adour) et Jérémy Ledoux (Hydrotube Énergie). Dans ce qui suit, nous résumons nos principales contributions à chacun d'eux.

Parallélisation en temps de l'assimilation de données

Nous commençons au Chapitre 2 en proposant une procédure pour coupler des méthodes d'assimilation de données séquentielles avec des algorithmes de parallélisation en temps, qui consiste à diviser l'intervalle de temps non borné en sous-intervalles de même longueur (*fenêtres*) et à appliquer ensuite, selon un ordre séquentiel, le solveur parallèle en temps sur chacun d'eux. En considérant l'observateur de Luenberger comme méthode d'assimilation, nous fournissons un critère d'arrêt qui préserve son taux exponentiel de convergence, ce qui donne une estimation a posteriori de la précision du solveur.

Afin d'aller plus loin, nous avons étudié plus précisément le cas d'une parallélisation en temps par l'algorithme Pararéel comme solveur parallèle en temps. Cela nous permet de concevoir un algorithme alternatif qui fournit une estimation a priori du nombre d'itérations nécessaires sur chaque fenêtre, ce qui nous permet également d'étudier l'efficacité théorique de l'ensemble de la procédure. Ces résultats sont basés sur une nouvelle estimation de convergence que nous dérivons pour Pararéel lorsque le solveur grossier est contractif.

Optimisation d'une bathymétrie

Nous passons à l'étude au Chapitre 3 de la détermination d'une bathymétrie à partir d'un problème d'optimisation, où une reformulation de l'équation de Helmholtz agit comme une contrainte. Même si cette équation est limitée à la description d'ondes de faible amplitude, elle est souvent utilisée en ingénierie en raison de sa simplicité, qui conduit à des solutions explicites lorsqu'une bathymétrie plate est considérée. En levant cette hypothèse, on obtient une formulation différente dans laquelle cette variable joue le rôle d'un diffuseur.

Sous des hypothèses appropriées sur la fonctionnelle de coût et l'ensemble admissible des bathymétries, nous sommes en mesure de prouver la continuité de la fonction contrôle-état et l'existence d'une solution optimale, en plus de la continuité et et le caractère borné de la vague résultante. Le problème de l'optimisation discrète est également abordé, en étudiant la convergence vers la solution optimale discrète ainsi que la convergence d'une approximation par éléments finis.

Analyse mathématique de la méthode de l'élément de pale

Nous abordons enfin au Chapitre 4 l'existence de solutions et la convergence des procédures de résolution pour la méthode de l'élément de pale. Le point clé de notre travail consiste à montrer que la décomposition proposée par Glauert peut être utilisée pour reformuler son ensemble original d'équations en une seule expression contenant deux termes bien distincts : un terme universel, indépendant de la turbine considérée et associé aux aspects macroscopiques du modèle ; et un terme expérimental, concernant les caractéristiques de la turbine et associé à la partie locale du modèle.

L'avantage de notre approche est qu'elle identifie explicitement les hypothèses sur les paramètres de la turbine qui garantissent l'existence d'une solution. De plus, elle nous aide aussi à présenter des critères de convergence pour différents algorithmes de résolution. Le modèle de Glauert est également utilisé pour optimiser la géométrie des pales, en ce sens qu'il maximise le rendement de la turbine. Nous rappelons les détails de la procédure de conception habituelle et discutons brièvement du cas où une correction du modèle est introduite.

État de l'art

Cette section est consacrée à un aperçu des différents sujets abordés dans cette thèse. Nous commençons par deux sujets étudiés dans le Chapitre 2 : les méthodes d'assimilation des données, en continu et en discret suivi d'un résumé des méthodes de parallélisation en espace et en temps, en particulier l'algorithme Pararéel. Nous passons ensuite à la discussion des éléments clés du Chapitre 3, l'estimation de la bathymétrie et modélisation des ondes, en mettant l'accent sur les principales hypothèses nécessaires pour dériver différents modèles de propagation des ondes. Enfin, le Chapitre 4 donne une brève description de la méthode de l'élément de pale.

Assimilation des données (AD)

Les modèles mathématiques sont largement utilisés pour décrire des systèmes complexes. Ils reposent sur des approximations et des simplifications d'un phénomène réel, ce qui définit en dernière instance son champ d'application. Après une validation appropriée, le but est de les utiliser pour décrire ou même prévoir une situation réelle, comme la surveillance sismique, la prévision de la grippe saisonnière ou l'estimation de l'état de charge d'une batterie [52]. Pour ce faire, nous devons intégrer des données réelles dans notre cadre. Les différentes techniques qui combinent les modèles mathématiques avec les observations disponibles pour améliorer la connaissance d'un système sont connues sous le nom d'*Assimilation des données* (AD). Parmi ces approches, nous rappelons ici les méthodes *séquentielles* et *variationnelles*.

Méthodes séquentielles

Supposons que pour une raison quelconque, par exemple des contraintes physiques ou budgétaires, nous n'ayons accès qu'à une information partielle sur l'état du système. Par exemple, dans certains cas, la condition initiale n'est pas connue avec précision, comme c'est le cas en science du climat [90]. Nous pouvons traiter ce manque d'information et cette incertitude en construisant un nouveau système qui utilise les observations disponibles pour se rapprocher de l'état réel. Dans un contexte déterministe, ce dispositif est appelé un *observateur*.

Nous supposons que le système est régi par

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \quad t \in [0, +\infty) \\ x(0) = x_0, \\ y(t) = Cx(t), \end{cases} \tag{1}$$

où $x \in \mathbb{R}^m$ est le vecteur d'état réel, $y \in \mathbb{R}^q$ représente les observations (avec $q < m$, $m, q \in \mathbb{N}^*$), u est une entrée et $x(0) = x_0$ une condition initiale inconnue. Les matrices A, B et C sont connues et leurs dimensions sont cohérentes.

L'*observateur de Luenberger* [65] imite le modèle précédent, mais il inclut un terme supplémentaire dans la dynamique qui mesure l'écart entre les observations et ses propres prédictions. Il produit une estimation d'état $\hat{x}$ satisfaisant

$$\begin{cases} \dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) + L\left[y(t) - \hat{y}(t)\right] & t \in [0, +\infty) \\ \hat{x}(0) = \hat{x}_0, \\ \hat{y}(t) = C\hat{x}(t), \end{cases}$$

avec $\hat{x}_0$ une condition initiale arbitraire. Tant que le modèle original (1) est observable (c-à-d que l'état initial $x(0)$ peut être déterminé uniquement à partir des observations dans $[0, T]$, pour tout T), l'erreur résultante $\|x(t) - \hat{x}(t)\|$ peut être mise à zéro à un taux exponentiel en choisissant correctement la *matrice de gain* L, puis l'état réel est récupéré asymptotiquement.

Une autre alternative est le *filtre de Kalman* [56], qui prend en compte les erreurs de mesure et modélise les incertitudes représentées par des bruits blancs gaussiens (à la fois stationnaires et mutuellement non corrélés), afin de construire une estimation d'état qui minimise l'erreur quadratique moyenne. Notons que des extensions au cas non linéaire ont été développées, e.g. le *observateur non linéaire de Luenberger* [3] et le *filtre de Kalman étendu* [53].

Une technique plus récente, développée par Auroux et Blum [6] est le *Nudging direct et rétrograde (Back and Forth Nudging, BFN)*. Le Nudging direct consiste simplement à ajouter un terme de rétroaction dans l'équation gouvernante, comme le fait l'observateur de Luenberger, mais aussi à supposer des observations complètes dans le (1) (c-à-d $y(t) = Cx_{obs}(t)$, avec C inversible) et aucune entrée $u(t)$. En utilisant les mêmes idées, le *Nudging rétrograde* considère un intervalle de temps borné $[0, T]$ pour approcher plutôt la condition initiale x_0 en résolvant

$$\begin{cases} \dot{\tilde{x}}(t) = A\tilde{x}(t) - K\left[x_{obs}(t) - \tilde{x}_k(t)\right] & \text{dans } [0, T] \\ \tilde{x}(T) = \tilde{x}_T \end{cases}$$

où $\tilde{x}_T$ est une observation du système à l'instant T et K est la *matrice de nudging*, choisie symétrique et positive définie pour assurer la convergence asymptotique

$$(\forall t \in (0, T]) \qquad \lim_{\substack{\min \\ \lambda \in \sigma(K)} \lambda \to +\infty} \tilde{x}(t) = x_{obs}(t). \tag{2}$$

L'algorithme BFN combine ces procédures en définissant la méthode itérative

$$\begin{cases} \dot{\hat{x}}_k(t) = A\hat{x}_k(t) + K\left[x_{obs}(t) - \hat{x}_k(t)\right] & \text{dans } [0, T] \\ \hat{x}_k(0) = \tilde{x}_{k-1}(0) \end{cases}$$

$$\begin{cases} \dot{\tilde{x}}_k(t) = A\tilde{x}_k(t) - K\left[x_{obs}(t) - \tilde{x}_k(t)\right] & \text{dans } [0, T] \\ \tilde{x}_k(T) = \hat{x}_k(T) \end{cases}$$

avec $\hat{x}_0(0) = x_0$. Sous certaines hypothèses, les deux suites $\{\hat{x}_k(t)\}_{k\geq 1}$ et $\{\tilde{x}_k(t)\}_{k\geq 1}$ convergent vers $\hat{x}_\infty(t)$ et $\tilde{x}_\infty(t)$, respectivement. De plus, ces fonctions limites présentent également le comportement asymptotique décrit dans (2), même pour $t = 0$.

Méthodes variationnelles

D'autre part, Sasaki [81] a proposé d'appliquer une approche différente aux problèmes de météorologie : ici, l'équation gouvernante contraint une fonctionnelle de coût J, représentant l'écart entre l'état réel $x(t)$ et les données disponibles. L'objectif de cette formulation est de minimiser $J = J(u)$, où $u(t)$ agit comme une entrée de l'équation gouvernante (par exemple, la condition initiale ou limite). Autrement dit, le fait de trouver la variable de contrôle u donne l'état réel $x = x(u)$. Les problèmes de ce type relèvent de la théorie du contrôle optimal appliquée aux EDP, dont le fondement théorique a été initialement développé par J.-L. Lions [60].

À titre indicatif, nous considérons le problème de minimisation continue

$$\begin{aligned} \min_{u\in\mathcal{U}} \quad & J(u) = \frac{1}{2}\int_\Omega (u - x_0^b)^\top B^{-1}(u - x_0^b)dx \\ & \qquad + \frac{1}{2}\int_0^T \int_\Omega (\mathbb{H}(x) - y)^\top R^{-1}(\mathbb{H}(x) - y)\, dxdt \\ \text{s.t.} \quad & \begin{cases} \dot{x}(t) = \mathbb{M}(x(t), t) & \text{dans } \Omega \times [0, T] \\ x(0) = u & \text{dans } \Omega \end{cases} \end{aligned} \tag{3}$$

La condition initiale u appartient à un espace fonctionnel $\mathcal{U}$ qui résume les propriétés souhaitables du contrôle, tandis que la définition de $J(u)$ fait intervenir plusieurs éléments : les matrices de covariance des erreurs de prévision et d'observation B et R, qui mesurent l'incertitude autour de l'estimation préalable de la condition initiale x_0^b et des observations $y(t)$, respectivement ; et un opérateur différentiel $\mathbb{H}$ qui décrit les observations prévues du système de contrôle. Ce dernier est également modélisé par un opérateur différentiel $\mathbb{M}$.

Nous avons besoin de calculer $\nabla J(u)$, soit pour dériver les conditions d'optimalité pour (3), soit pour résoudre numériquement ce problème. Dans ce qui suit, nous décrivons différentes approches pour le faire [4, 74, 53].

Un calcul direct. Puisque x dépend implicitement de u, nous devons d'abord déterminer son comportement lorsque la condition initiale est légèrement perturbée dans une direction v. Cette variation, désignée par $\mathcal{X}$, satisfait

$$\begin{cases} \dot{\mathcal{X}}(t) = \dfrac{\partial \mathbb{M}}{\partial x}(x(t), t)\mathcal{X} & \text{dans } \Omega \times [0, T] \\ \mathcal{X}(0) = v & \text{dans } \Omega \end{cases} \tag{4}$$

et ensuite, un calcul direct donne

$$\langle \nabla J(u), v \rangle = \int_\Omega \Big[B^{-1}(u - x_0^b) + \int_0^T \mathcal{X}^\top \Big(\frac{\partial \mathbb{H}}{\partial x}(x(t), t) \Big)^* R^{-1}(\mathbb{H}(x) - y)\, dt \Big] v\, dx.$$

où * désigne l'opérateur adjoint. L'inconvénient de cette méthode est qu'elle nécessite de résoudre (4) pour chaque direction v.

La méthode adjointe. Cette procédure réduit (3) à un problème d'optimisation sans contrainte avec des variables supplémentaires, en définissant le Lagrangien

$$\mathcal{L}(u, x, \lambda, \mu) = J(u) + \int_0^T \int_\Omega \lambda^\top \left[\dot{x}(t) - \mathbb{M}(x(t), t) \right] dxdt + \int_\Omega \mu^\top (x(0) - u)\, dx,$$

où λ et μ sont connus sous le nom des multiplicateurs de Lagrange associés à chaque équation du système principal.

Au lieu de traiter l'équation linéaire tangente (4), la fonction Lagrangienne nous permet de dériver une équation pour la variable duale de $\mathcal{X}$ de la façon suivante. La mise à zéro de la dérivée du Lagrangien par rapport à x dans la direction z, à (u, x, λ, μ), conduit à

$$\langle \nabla_x \mathcal{L}(u, x, \lambda, \mu), z \rangle = 0$$
$$\int_0^T \int_\Omega z^\top \Big(\frac{\partial \mathbb{H}}{\partial x}(x(t), t) \Big)^* R^{-1}(\mathbb{H}(x) - y)\, dxdt$$
$$\int_0^T \int_\Omega \lambda^\top \left[\dot{z}(t) - \frac{\partial \mathbb{M}}{\partial x}(x(t), t) z \right] dxdt + \int_\Omega \mu^\top z(0)\, dx = 0,$$

et après avoir intégré par parties la deuxième intégrale et en réorganisant les termes, nous obtenons

$$\int_0^T \int_\Omega z^\top \Big[\Big(\frac{\partial \mathbb{H}}{\partial x}(x(t), t) \Big)^* R^{-1}(\mathbb{H}(x) - y) + \dot{\lambda} - \Big(\frac{\partial \mathbb{M}}{\partial x}(x(t), t) \Big)^* \lambda \Big] dxdt$$
$$+ \int_\Omega \lambda^\top(T) z(T) + \int_\Omega [\mu - \lambda(0)]^\top z(0)\, dx = 0.$$

Puisque z est une direction arbitraire, λ doit nécessairement satisfaire l'*équation adjointe*

$$\begin{cases} \dot{\lambda}(t) - \Big(\frac{\partial \mathbb{M}}{\partial x}(x(t), t) \Big)^* \lambda = - \Big(\frac{\partial \mathbb{H}}{\partial x}(x(t), t) \Big)^* R^{-1}(\mathbb{H}(x) - y) & \text{in } \Omega \times [0, T] \\ \lambda(T) = 0 & \text{in } \Omega \end{cases} \tag{5}$$

et la condition supplémentaire $\mu^\top = \lambda(0)$. Ensuite, nous pouvons calculer le gradient pour n'importe quelle direction v par

$$\langle \nabla J(u), v \rangle = \langle \nabla_u \mathcal{L}(u, x, \lambda, \mu), v \rangle = \int_\Omega \Big[B^{-1}(u - x_0^b) - \lambda^\top(0) \Big] v\, dx.$$

Notons que nous devons résoudre (5) une seule fois, puisque la variable adjointe ne dépend pas de v.

Méthodes variationnelles discrètes. Contrairement au cadre continu, qui nécessite de résoudre l'équation linéaire tangente (4) ou l'équation adjointe pour obtenir le gradient, la discrétisation (3) conduit à son calcul direct. Ensuite, les méthodes variationnelles discrètes diffèrent dans la façon dont elles décrivent l'équation gouvernante et le traitement des non-linéarités possibles.

Étant donné une discrétisation $\{t_n\}_{n=0}^N$ de $[0,T]$, nous désignons par x_i et y_i l'état et le vecteur d'observation au temps t_i, respectivement. En utilisant la même notation qu'auparavant sur les variables J, u, x_0^b, B et R, même si elles peuvent dépendre de la discrétisation temporelle considérée, l'algorithme *4D-Var* se lit ainsi

$$\begin{aligned} &\min_{u\in\mathcal{U}} J(u) = \frac{1}{2}(u-x_0^b)^\top B^{-1}(u-x_0^b) + \frac{1}{2}\sum_{n=1}^{N}(\mathcal{H}(x_n)-y_n)^\top R^{-1}(\mathcal{H}(x_n)-y_n) \\ &\text{s.t.} \begin{cases} x_n = \mathcal{M}_{[t_{n-1},t_n]}(x_{n-1}) & \forall n = 1,\dots,N \\ x_0 = u \end{cases} \end{aligned} \tag{6}$$

où $\mathcal{H}$ est l'opérateur d'observation et $\{\mathcal{M}_{[t_{n-1},t_n]}(\cdot)\}_{n=1}^N$ est une famille d'opérateurs qui décrivent les transitions de l'état de t_{n-1} à t_n. Puis

$$x_n = \left(\mathcal{M}_{[t_{n-1},t_n]} \circ \cdots \circ \mathcal{M}_{[t_0,t_1]}\right)(u) := \mathcal{M}_{[t_0,t_n]}(u)$$

et le gradient peut être calculé par

$$\nabla J(u) = B^{-1}(u-x_0^b) + \sum_{n=1}^{N} \boldsymbol{M}_n^\top \boldsymbol{H}_n^\top \cdot R^{-1}(\mathcal{H}\circ\mathcal{M}_{[t_0,t_n]}(u) - y_n). \tag{7}$$

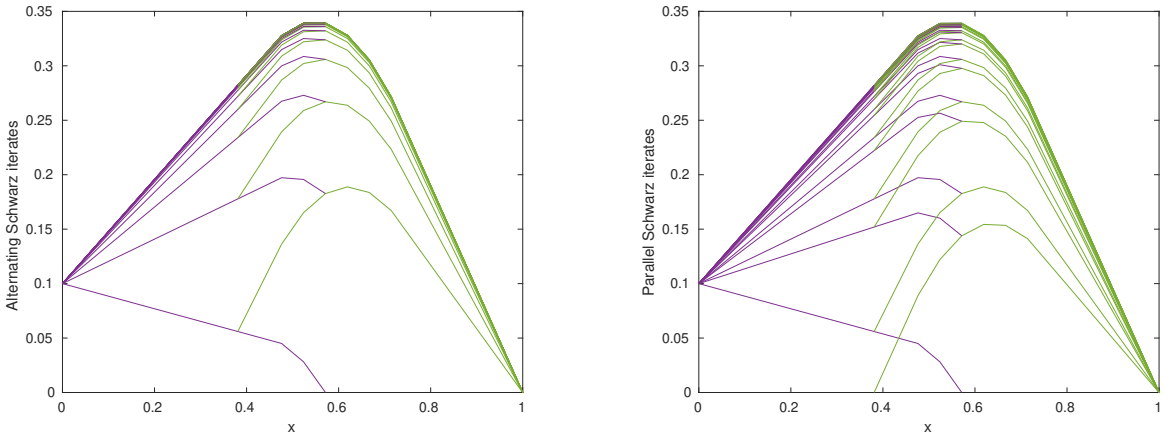
Les matrices $\boldsymbol{M}_n = D\mathcal{M}_{[t_0,t_n]}(u)$ et $\boldsymbol{H}_n = D\mathcal{H}(x_n)$ sont connues comme les opérateurs linéaires tangentes associés à $\mathcal{M}_{[t_0,t_n]}$ et $\mathcal{H}$, respectivement.

D'autres variantes de cette méthode sont *3D-Var*, une version indépendante du temps qui peut aussi être appliquée à des problèmes évolutifs en supposant que toutes les observations ne sont disponibles qu'au début (c-à-d qu'aucun système principale n'est nécessaire) ; *3D-FGAT*, une amélioration de la dernière dans laquelle nous remplaçons seulement $\boldsymbol{M}_n$ par la matrice d'identité, simplifiant le calcul du gradient ; et *4D-Var Incrémental* [24], qui consiste à approximer (6) en utilisant une séquence de problèmes de minimisation quadratique pour réduire le coût opérationnel.

Méthodes parallèles en espace-temps

Pour l'une ou l'autre des méthodes d'AD susmentionnées, le calcul numérique de l'estimation de l'état est dans de nombreuses situations aussi pertinent que sa précision, c'est-à-dire que le problème est posé sous contrainte de calcul en temps réel. Ainsi, la première exige souvent d'être réalisée dans un délai raisonnable, ce qui est possible à l'aide de méthodes parallèles dans l'espace ou dans le temps, une approche naturelle pour accélérer la résolution numérique des EDP à l'aide du calcul parallèle. En suivant Gander [33, 34], nous en décrivons brièvement quelques-unes.

Au cours du dix-neuvième siècle, l'analyse de Fourier a été le principal outil d'étude des EDP, bien qu'elle soit limitée à des géométries simples comme les cercles ou les rectangles. Dans le but d'étendre le principe de Dirichlet à des domaines arbitraires, Schwarz [82] a proposé de résoudre l'équation de Laplace en décomposant le domaine en deux sous-domaines qui se chevauchent, où un calcul explicite peut-être utilisé, par exemple via l'analyse de Fourier, puis en résolvant alternativement un problème réduit sur chacun, dans une procédure connue aujourd'hui sous le nom de *Méthode de Schwarz alternée*. Cette décomposition est le principe sous-jacent des méthodes de *décomposition de domaines*.



(a) Algorithme alterné (b) Algorithme parallèle

Figure 1: Méthode de Schwarz (discrétisée) appliqué à $-\frac{d^2u}{dx^2} + \eta u = f$ in $[0,1]$ [37, p.8].

Cent ans plus tard, Lions [61, 62, 63] a étendu la méthode précédente à un cadre parallèle, en envisageant la possibilité d'absence de chevauchement des sous-domaines et en résolvant de manière synchrone les problèmes locaux associés. La *Méthode de Schwarz parallèle* était née. Depuis lors, à une époque où les ordinateurs deviennent de plus en plus performants, de nombreuses méthodes ont été développées pour tirer parti de cette stratégie. Mais qu'en est-il de la dimension temporelle ? Comme la solution d'une EDP évolutive est naturellement affectée par le passé, le temps n'est généralement pas utilisé dans le calcul parallèle, même si des méthodes parallèles en temps ont été développées depuis plus de 50 ans. Ses origines remontent à Nievergelt [72], qui a proposé l'idée principale derrière les *Méthodes de tir multiples* : décomposer l'intervalle de temps en sous-intervalles disjoints et résoudre simultanément une famille de problèmes de valeur initiale, en brisant la nature séquentielle intrinsèque de l'équation différentielle dépendante du temps.

D'une manière plus générale, les méthodes de parallélisation espace-temps se distinguent par le caractère itératif ou direct de la procédure et la décomposition du domaine spatio-temporel considéré. Divisées en quatre classes, les méthodes de *Tir multiples* parallélisent sur l'intervalle de temps, les méthodes de *relaxation d'ondes et décomposition de domaine* utilisent plutôt la dimension spatiale et les méthodes *Multigrille* travaillent simultanément sur les deux, étant toutes de nature itérative. En revanche, les méthodes *parallèles à temps direct* tentent de récupérer une solution en utilisant un solveur direct.

Dans ce qui suit, nous rappelons les bases de l'algorithme Parareal, une des plus récentes méthodes de tir multiple ; et quelques exemples de parallélisation de problèmes d'assimilation de données.

L'algorithme Pararéel

La parallélisation en temps du problème

$$\begin{cases} \dot{u}(t) = f(u(t)), \quad t \in [0, T] \\ u(0) = u_0 \end{cases}$$

exige de décomposer l'intervalle de temps en N sous-intervalles, désignés par (t_{n-1}, t_n); et d'introduire des cibles intermédiaires qui servent de conditions initiales sur chacune. Une façon directe de déterminer ces valeurs consiste à résoudre un système d'équations non linéaire en appliquant la méthode de Newton, une procédure coûteuse par un système de grande taille puisqu'elle repose sur le calcul d'une matrice jacobienne.

Lions, Maday et Turinici [64] ont proposé l'algorithme Pararéel, une approche différente qui utilise deux solveurs $\mathcal{F}$ et $\mathcal{G}$ qui calculent une approximation numérique fine et grossière de u sur les sous-intervalles et mettent à jour les conditions initiales artificielles. Gander et Vandewalle [39] ont prouvé que cette méthode se lit comme une méthode de tir multiple dans laquelle la matrice jacobienne est approximée par une différence finie dans une grille grossière.

L'algorithme Pararéel s'approche de $\{u(t_n)\}_{n=1}^N$ par une séquence $\{U_n^k\}_{n=1}^N$, qui est construite comme suit :

(a) puisque les conditions initiales sont inconnues sauf sur le premier sous-intervalle, nous imposons des valeurs arbitraires sur le reste en utilisant le solveur grossier $\mathcal{G}$,

$$\begin{aligned} U_n^0 &= \mathcal{G}(t_n, t_{n-1}, U_{n-1}^0), \\ U_0^0 &= u_0. \end{aligned}$$

où $\mathcal{G}(t_n, t_{n-1}, U_{n-1}^0)$ désigne la solution obtenue avec le solveur grossier à t_n, en considérant U_{n-1}^0 comme condition initiale à t_{n-1}.

(b) Puis résoudre en parallèle les problèmes restreints

$$\begin{cases} \dot{u}(t) = f(u(t)), & t \in [t_{n-1}, t_n] \\ u(t_{n-1}) = U_{n-1}^k \end{cases}$$

en utilisant le solveur fin $\mathcal{F}$, ce qui donne les approximations $\{\mathcal{F}(t_n, t_{n-1}, U_{n-1}^k)\}_{n=1}^N$.

(c) Enfin, lissez les discontinuités précédemment introduites en définissant la suite

$$U_n^{k+1} := \mathcal{F}(t_n, t_{n-1}, U_{n-1}^k) + \mathcal{G}(t_n, t_{n-1}, U_{n-1}^{k+1}) - \mathcal{G}(t_n, t_{n-1}, U_{n-1}^k),$$

où l'exposant k indique le nombre d'itérations en cours. Notons que sur le côté droit, le premier et le troisième terme ont déjà été calculés, alors que le second montre que la mise à jour doit être faite de façon séquentielle. Heureusement, ce n'est pas coûteux en calcul car cela ne dépend que du solveur grossier $\mathcal{G}$.

L'avantage de cet algorithme est son taux de convergence superlinéaire. En effet, grâce à Gander et Hairer, nous avons l'estimation a priori :

Theorem (Convergence du Pararéel [36])**.** *Soit $\mathcal{F}(t_n, t_{n-1}, U_n^k)$ la solution exacte sur le sous-domaine temporel (t_{n-1}, t_n) et soit $\mathcal{G}(t_n, t_{n-1}, U_n^k)$ une solution approximative avec une erreur de troncature locale bornée par $C_3\Delta T^{p+1}$, et satisfaisant*

$$\mathcal{F}(t_n, t_{n-1}, x) - \mathcal{G}(t_n, t_{n-1}, x) = c_{p+1}(x)\Delta T^{p+1} + c_{p+2}(x)\Delta T^{p+2} + \dots,$$

pour ΔT petit, où les coefficients $\{c_j\}_{j\geq p+1}$ sont continuellement différentiables, et supposons que $\mathcal{G}$ satisfait la condition de Lipschitz

$$\|\mathcal{G}(t + \Delta T, t, x) - \mathcal{G}(t + \Delta T, t, y)\| \leq (1 + C_2\Delta T)\|x - y\|. \tag{8}$$

Ensuite, à l'itération k de l'algorithme Pararéel, nous avons la borne

$$\left\|u(t_n) - U_n^k\right\| \leq \frac{C_3}{C_1}\frac{(C_1\Delta T^{p+1})^{k+1}}{k!}(1 + C_2\Delta T)^{n-(k+1)}\prod_{j=0}^{k}(n-j). \tag{9}$$

En raison du terme produit dans (9), après k itérations de l'algorithme Parareal, l'approximation est exacte sur les premiers k sous-intervalles (comme le montre la figure 2), et donc elle converge au plus en N itérations.

Même si ce n'est pas explicitement indiqué, la constante C_2 doit être positive, raison pour laquelle l'hypothèse de Lipschitz (8) ne prend pas en compte le cas décroissant, c-à-d lorsque le solveur grossier est contractif. Puisque nous sommes intéressés à coupler cet algorithme avec l'observateur de Luenberger et à tirer profit de son comportement de décroissance, nous présentons dans le Chapitre 2 une variante du Théorème précédent qui couvre cette situation.

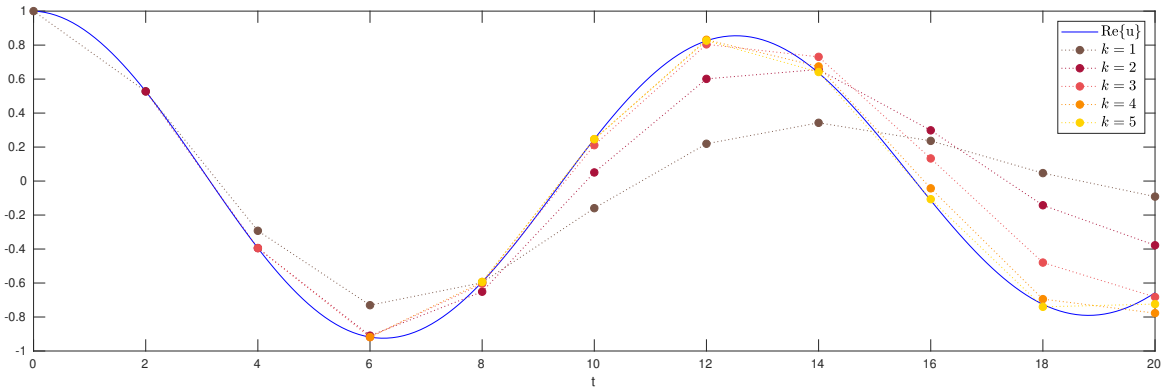


Figure 2: L'algorithme Pararéel appliqué à l'equation de Dahlquist $\dot{u}(t) = -\frac{\mathrm{i}}{2}u$ in $[0, 20]$

Méthodes de parallélisation en espace-temps et l'AD

Trémolet et Le Dimet [88] ont été parmi les premiers à aborder la parallélisation des problèmes d'assimilation de données variationnelles en météorologie. Dans un cadre continu, ils ont proposé une approche de décomposition des domaines combinée avec la méthode adjointe, en assignant à chaque sous-domaine une version locale de (3), avec un terme supplémentaire sur la fonctionnelle de coût locale pour renforcer la continuité de l'état entre les domaines adjacents.

Vingt ans plus tard, des nouvelles stratégies dans ce domaine incluent également la direction temporelle. Par exemple, Rao et Sandu [79] appliquent un solveur quasi-Newton au problème 4D-Var, mais ils parallélisent en temps le calcul du gradient. Une approche plus sophistiquée est proposée par D'Amore et Cacciapuoti [26], qui combine l'algorithme Pararéel avec la méthode de Schwarz multiplicatif (MPS) pour résoudre 4D-Var.

La méthodologie associée à l'algorithme Parareal a également été appliquée à des problèmes d'optimisation, mais pas nécessairement liés à l'AD. Par exemple, Maday, Salomon et Turinici [66] ont proposé un algorithme spécifique pour résoudre les systèmes d'optimalité dans le cas du contrôle quantique. En définissant des états intermédiaires et en résolvant ensuite en parallèle une famille de problèmes d'optimisation locaux, ces valeurs sont mises à jour après avoir résolu séquentiellement les équations gouvernantes et adjointes, ce qui est peu coûteux en termes de calcul par rapport à la procédure d'optimisation. Comme la fonctionnelle de coût original n'est pas parallélisable, la méthode en utilise une autre qui dépend de la variable adjointe et peut également être décomposée comme la somme des fonctionnelles de coût pour chaque sous-intervalle. En définitive, sa solution optimale coïncide avec celle du problème original.

Estimation de la bathymétrie

Récemment, une littérature considérable s'est développée autour du sujet des problèmes inverses dans les écoulements de surface libre. Une revue de Sellier identifie différentes techniques appliquées à la reconstruction bathymétrique [83, Section 4.2], qui reposent principalement sur la dérivation d'une formule explicite pour la bathymétrie, la résolution numérique des équations gouvernantes, ou l'assimilation des données variationnelles [51].

Une approche variationnelle est également bien adaptée à la résolution de problèmes d'ingénierie côtière, car ils doivent satisfaire plusieurs contraintes mécaniques. Par exemple, parmi les nombreux aspects à considérer lors de la conception d'un port, la construction de structures de protection est essentielle pour le protéger contre l'impact des vagues. Ces structures peuvent être optimisées pour minimiser l'énergie des vagues, en étudiant son interaction avec les vagues réfléchies [55]. Bouharguane et Mohammadi [69, 14] envisagent une approche temporelle pour étudier l'évolution du mouvement du sable au fond de la mer, qui pourrait également permettre à ces structures de changer dans le temps. Dans ce cas, les fonctionnelles proposées sont minimisées localement en utilisant l'analyse de sensibilité, une technique largement appliquée dans les géosciences.

D'un point de vue mathématique, la résolution de ce genre de problème est surtout numérique. Les questions concernant un espace de contrôle approprié pour la bathymétrie, la contrôlabilité, la régularité de la solution ou le caractère bien posé du problème ne sont généralement pas abordées. Une approche théorique appliquée à la modélisation des bassins de surf peut être trouvée dans les publications [25, 71], où le but est de maximiser l'énergie totale de la vague prescrite. La première propose de déterminer une bathymétrie, tandis que la seconde définit la forme et le déplacement d'un objet sous-marin à une profondeur constante.

Modélisation des vagues

Dans tous les problèmes précédents, une contrainte cruciale pour déterminer la bathymétrie est son interaction avec les vagues. La pierre angulaire de sa modélisation sont les équations de Navier-Stokes

$$\begin{cases} \rho\left(\dfrac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\,\mathbf{u}\right) = \operatorname{div}(\sigma_T) + \mathbf{g} & \text{dans } \Omega_t, \\ \operatorname{div}(\mathbf{u}) = 0 & \text{dans } \Omega_t, \\ \mathbf{u} = \mathbf{u}_0 & \text{dans } \Omega_0, \end{cases}$$

où $\mathbf{u} = (u, v, w)^\top$ représente la vitesse du fluide dans la région bornée et dépendante du temps Ω_t ; ρ, $\mathbf{g}$ et σ_T représentent respectivement le coefficient de densité, la gravité et le tenseur de contrainte. Nous omettons ici les conditions aux limites au niveau

de l'eau $\eta(x,t)$ et au fond $-z_b(x)$, éventuellement en fonction des effets de la pression, de la viscosité et du frottement, voir la section 3.1.1 pour plus de détails. Dans ce qui suit, par souci de simplicité, nous négligeons les deux derniers.

Certaines simplifications des équations de Navier-Stokes donnent lieu à différents modèles de propagation des vagues [16, 59], tous dépendant de la relation entre trois paramètres caractéristiques appelés profondeur relative H, longueur horizontale L et amplitude verticale maximale A. Pour rappeler certains d'entre eux, nous introduisons les rapports

$$\varepsilon = \frac{H}{L},\ \delta = \frac{A}{H}.$$

Les rivières, les domaines côtiers et les océans peuvent être modélisés via une intégration verticale en utilisant l'hypothèse d'eau peu profonde $\varepsilon \ll 1$, alors que les vagues de petite amplitude sont décrites par des modèles où $\delta \ll 1$.

En supposant un régime d'eau peu profonde, ainsi que différentes valeurs de δ et des approximations de la pression, on obtient les modèles suivants.

Équations de Saint-Venant. La *pression hydrostatique*, c-à-d que la contribution de l'accélération verticale du fluide dans la pression est négligeable, combinée avec $\delta = \mathcal{O}(1)$ donne

$$\begin{cases} \dfrac{\partial \eta}{\partial t} + \dfrac{\partial h\overline{u}}{\partial x} + \dfrac{\partial h\overline{v}}{\partial y} = 0 \\ \dfrac{\partial h\overline{u}}{\partial t} + \dfrac{\partial h\overline{u}^2}{\partial x} + \dfrac{\partial h\overline{u}\,\overline{v}}{\partial y} + gh\dfrac{\partial \eta}{\partial x} = 0 \\ \dfrac{\partial h\overline{v}}{\partial t} + \dfrac{\partial h\overline{u}\,\overline{v}}{\partial x} + \dfrac{\partial h\overline{v}^2}{\partial x} + gh\dfrac{\partial \eta}{\partial y} = 0 \end{cases} \tag{10}$$

où $h = \eta + z_b$ est la hauteur d'eau et $\overline{u}$, $\overline{v}$ sont les vitesses moyennes sur la profondeur. Ce modèle est couramment utilisé pour modéliser les ruptures de barrage [40], les sauts hydrauliques et les flux de marée dans les estuaires.

Équation des ondes. Maintenir la pression hydrostatique mais en supposant que $\delta \ll 1$ est équivalent à négliger les termes d'accélération convective dans (10) et linéariser le nouveau système autour du niveau de la mer $\eta = 0$. En combinant les expressions résultantes, on obtient

$$\frac{\partial^2 \eta}{\partial t^2} - \nabla\left(g z_b \nabla \eta\right) = 0.$$

En particulier, sa solution a pour forme $\eta(x,t) = \mathrm{Re}\{\psi_{tot}(x)e^{-i\omega t}\}$, où ω est une fréquence donnée et l'amplitude ψ_{tot} satisfait l'*équation de Helmholtz*

$$\triangle \psi_{tot} + \left(\frac{\omega^2}{g z_b}\right)\psi_{tot} = 0,$$

quand la profondeur $-z_b$ est supposée constante. Dans des conditions aux limites appropriées, cette équation est utilisée pour étudier les problèmes de diffusion des vagues, comme celle présentée au Chapitre 3, où une bathymétrie variable agit comme un diffuseur.

Équations de Boussinesq classique. D'autre part, une pression non hydrostatique introduit des termes dispersifs dans les équations gouvernantes. Un troisième rapport, appelé le nombre d'Ursell

$$U_r = \frac{\delta}{\varepsilon^2},$$

mesure leur force par rapport aux termes non linéaires [89]. En mouvement bidimensionnel, un choix spécifique de ce paramètre ($U_r = \mathcal{O}(1)$ et $\delta \ll 1$) conduit à

$$\begin{cases} \dfrac{\partial \eta}{\partial t} + \dfrac{\partial h\overline{u}}{\partial x} = 0 \\ \dfrac{\partial \overline{u}}{\partial t} + \overline{u}\dfrac{\partial \overline{u}}{\partial x} + g\dfrac{\partial \eta}{\partial x} = \dfrac{z_b^2}{3}\dfrac{\partial^3 \overline{u}}{\partial x^2 \partial t} \end{cases}$$

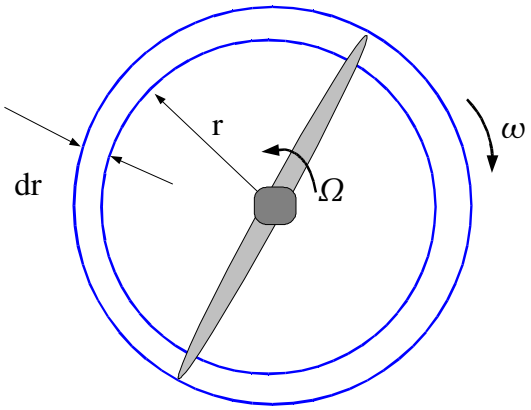
pour une profondeur constante $-z_b$. Ce modèle et ses variantes ultérieures sont utilisés pour décrire les mouvements de flottabilité dans les fluides, comme les vagues entrant dans la zone proche du rivage [10] ou les vagues générées par les glissements de terrain [28].

Méthode de l'élément de pale

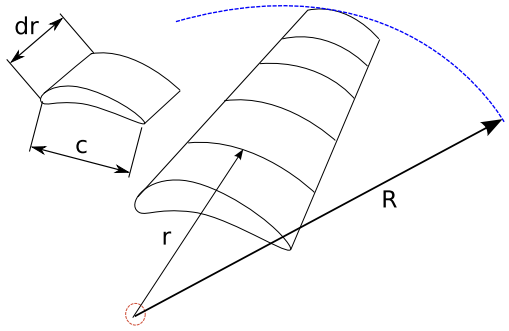
La *méthode de l'élément de pale* (Blade element momentum theory, BEM) est un modèle mécanique largement utilisé pour évaluer la performance des turbines, en fonction des caractéristiques mécaniques et géométriques de leurs pales et du courant auquel elles sont exposées. Développé par Glauert [44], il résulte de la combinaison de deux modèles différents : *Blade element theory* (BET) et la *Théorie de Froude* (Momentum theory, MT).

En découpant la pale en sections qui sont traitées selon un modèle planaire, BET étudie le comportement de la turbine d'un point de vue local [32]. Les quantités fondamentales de ce modèle sont les coefficients C_L et C_D appelés *portance* et *traînée*, qui sont introduits pour tenir compte ces forces exprimées dans le plan de coupe. Les résultats sont ensuite intégrés le long de la pale pour obtenir les quantités d'intérêt global. En contraste, la Théorie de Froude [78] est une théorie globale qui étudie de façon macroscopique le comportement d'une colonne de fluide traversant une turbine.

La théorie BEM repose alors sur une décomposition du système fluide/turbine en une partie macroscopique via la MT et une partie planaire locale via la BET. Cette dernière considère une décomposition radiale pour les pales et la colonne de fluide (Figure 3a), en divisant la zone du rotor en anneaux concentriques d'épaisseur infinitésimale qui n'interagissent pas entre eux.



(a) Décomposition radiale



(b) Décomposition de la pale en éléments annulaires

Figure 3: Décompositions impliquées dans la théorie BEM [54, p.8].

Le modèle de Glauert lie finalement trois variables associées à l'anneau considéré : le *facteur d'induction axiale* a, le *facteur d'induction tangentielle* a' et le *déviation de l'angle relatif* φ. Étant donné un profil de pale et en supposant une vitesse de rotation fixe, les équations résultantes sont

$$\tan\varphi = \frac{1-a}{\lambda(1+a)}, \tag{11}$$

$$\frac{a}{1-a} = \frac{1}{4\sin^2\varphi} \cdot \frac{Bc_\lambda}{2\pi r}\left(C_L(\varphi-\gamma_\lambda)\cos\varphi + C_D(\varphi-\gamma_\lambda)\sin\varphi\right), \tag{12}$$

$$\frac{a'}{1-a} = \frac{1}{4\lambda\sin^2\varphi} \cdot \frac{Bc_\lambda}{2\pi r}\left(C_L(\varphi-\gamma_\lambda)\sin\varphi - C_D(\varphi-\gamma_\lambda)\cos\varphi\right). \tag{13}$$

où r indique la distance entre l'élément de pale pris en compte (figure 3b) et le rotor, $\lambda = \lambda(r)$ est le rapport de vitesse au bout et B le nombre total de pales. Le *corde* c_λ et l'*angle de torsion* γ_λ sont donnés par la géométrie des pales.

Diverses modifications ont été apportées au modèle pour tenir compte, par exemple, de l'écoulement autour de l'extrémité d'une pale ou d'un état de sillage turbulent. Pour un compte rendu plus détaillé, voir la section 4.1.4.

Enfin, nous évaluons le rendement de la turbine en utilisant le *coefficient de puissance*

$$C_p(\varphi) = \frac{8}{\lambda_{\max}} \int_{\lambda_{\min}}^{\lambda_{\max}} \lambda^3 a'(1-a) \left(1 - \frac{C_D}{C_L}(\varphi - \gamma_\lambda) \tan^{-1} \varphi\right) d\lambda.$$

Inversement, nous pouvons concevoir un profil de pale qui maximise cette quantité en résolvant

$$\begin{aligned} \max_{\varphi} \quad & C_p(\varphi) \\ s.t \quad & (11\text{-}13) \\ & \overline{\alpha} = \varphi - \gamma_\lambda, \end{aligned}$$

où $\overline{\alpha}$ est choisi de telle sorte que le rapport $\frac{C_D}{C_L}$ soit proche de zéro. Afin de simplifier ce problème d'optimisation, on suppose généralement un coefficient de traînée nul, ce qui conduit à une solution explicite. A notre connaissance, le cas général concernant la traînée non nulle et les corrections possibles du modèle n'a pas été abordé.

General Introduction

The present thesis aims to contribute to the development of a theoretical framework for three problems in the context of renewable marine energy, namely the time-parallelization of sequential data assimilation problems, bathymetry optimization and mathematical analysis of the blade element momentum theory. Since their solving rely largely on empirical knowledge, we believe that adopting a mathematical standpoint leads to a better understanding of the different situations, this being an opportunity to encourage interdisciplinary collaboration between mathematics and applied sciences as geophysics and engineering. In what follows, we briefly introduce each of these topics.

Time-parallelization of sequential data assimilation problems

The assumptions behind a mathematical model not only determine their range of applicability, but also induce an inevitable gap between predictions and reality. In order to narrow this difference, we can sacrifice the simplicity of the model or incorporate real data instead, by following a Data assimilation (DA) procedure. Several aspects have to be considered when applying these techniques to e.g. meteorology or oceanography problems, but here we recall that due to the number of state variables and the vast amount of observations required, their numerical solving is computationally expensive. Trémolet and Le Dimet [88] were among the first to address the parallelization of *variational* DA problems (which are based on optimal control and use the information collected in a fixed amount of time) by using a Domain decomposition (DD) approach. Since then, the coupling between these two procedures have been widely studied.

Roughly speaking, DD methods consist in decomposing the spatial dimension into subdomains, with possible overlap, and then solve syncronously a local problem on each. This divide-and-conquer strategy seems counterintuitive when handling the time direction, due to its inherently sequential nature, which is why time-parallelization is not commonly applied to DA problems.

Parallel-in-time algorithms can be quite useful when dealing with large time intervals, as is the case of *sequential* DA methods, where information can arrive uninterrupted. Then a natural question arises: can we combine both procedures? We begin to answer this question in Chapter 2, by studying the Luenberger observer and its coupling with the Parareal algorithm.

Bathymetry optimization

Despite the fact that the bathymetry can be inaccurately known in many situations, wave propagation models strongly depend on this parameter to capture the flow behavior, which emphasize the importance of studying inverse problems concerning its reconstruction from observed free surface data.

These kinds of problem are usually solved by simply discretizing the governing equations or with the help of sequential DA methods. Another alternative is to consider the bathymetry as control variable of a PDE-constrained optimization problem, an approach used in coastal engineering due to mechanical constraints associated with building structures and their interaction with sea waves. However, its solving rely mostly in sensitivity analysis, numerical methods or simplifications of the model that leads to explicit solutions and then, questions regarding a suitable control space, continuity of the control-to-state mapping, regularity or well-posedness of the solution are in general not adressed. Recently, Dalphin and Barros [25] carried out this theoretical analysis to model a wavemaker.

In Chapter 3 we try to answer the questions above when considering a reformulation of the Helmholtz equation for modeling wave propagation and a general cost functional that can be identified, for instance, with the mismatch between the predicted and observed wave solution.

Mathematical analysis of the Blade element momentum theory

Hydrotube Énergie, enterprise devoted to nautical design and on-board electronics systems, tested in 2015 a propotype of a floating water turbine on the Garonne river, in Bordeaux. But moving towards industrial scale production requires the development of a numerical software to simulate an optimized device, which is the purpose of its partnership with team ANGE (INRIA), a research group working in modeling, analysis and simulation of geophysical flows.

Among the problems that need to be addressed, we study the optimization of the turbine efficiency via the *Blade element momentum (BEM) theory*. Introduced by Glauert [44], this method is widely used for determining the efficiency and hence the design parameters of a blade, according to their geometric and mechanical characteristics and the current to which it is exposed. It follows from the combination of two independent methods that treat the fluid/turbine system from a macroscopic and local perspective, which results in a set of equations that are solved by applying an iterative algorithm. In some cases, a correction factor is required to ensure existence of solutions of the model.

Although old, this method is still used due to its relative simplicity compared to the complexity of the hydrodynamic phenomenon developed in the fluid/turbine system. Nevertheless, both the existence of solutions and the algorithm convergence has been never analyzed from a mathematical point of view, being this the aim of Chapter 4.

Contributions of this thesis

This work is divided into three independent chapters regarding the aforementioned topics, in which I had the pleasure to collaborate with Julien Salomon (INRIA Paris), Felix Kwok (Hong Kong Baptist University), Pierre-Henri Cocquet (Université de Pau et des Pays de l'Adour) and Jérémy Ledoux (Hydrotube Énergie). In the following, we summarize our main contributions to each.

Time-parallelization of sequential data assimilation problems

We start in Chapter 2 by proposing a procedure to couple sequential data assimilation methods with parallel-in-time algorithms, that consists in splitting the unbounded time interval into subintervals of the same length (*windows*) and then apply, following a sequential order, the time-parallel solver on each. By considering the Luenberger observer as assimilation method, we provide a stopping criterion that preserves its exponential rate of convergence, which yields an a posteriori estimate of the accuracy of the solver.

In order to go further, we set the Parareal algorithm as parallel-in-time solver. This allows us to design an alternative algorithm that provides an a priori estimate of the number of iterations required on each window, which also enables us to investigate the theoretical efficiency of the entire procedure. These results are based on a new convergence estimate that we derive for Parareal when the coarse solver is contractive.

Bathymetry optimization

We move on to study in Chapter 3 the determination of a bathymetry from an optimization problem, where a reformulation of the Helmholtz equation acts as a constraint. Even though this equation is limited to describe waves of small amplitude, it is often used in engineering due to its simplicity, which leads to explicit solutions when a flat bathymetry is assumed. By lifting this hypothesis, we obtain a different formulation in which this variable plays the role of a scatterer.

Under suitable assumptions on the cost functional and the admissible set of bathymetries, we are able to prove the continuity of the control-to-state mapping and the existence of an optimal solution, in addition to the continuity and boundedness of the resulting wave. The discrete optimization problem is also addressed, studying

the convergence to the discrete optimal solution as well as the convergence of a finite element approximation.

Mathematical analysis of the Blade element momentum theory

We finally discuss in Chapter 4 the existence of solutions and convergence of solving procedures for the Blade element momentum theory. The key point of our work consists in showing that the decomposition proposed by Glauert can be used to reformulate its original set of equations into a single expression containing two very distinct terms: a universal one, independent of the turbine considered and associated with macroscopic aspects of the model; and an experimental term, concerning the characteristics of the turbine and related with the local part of the model.

The advantage of our approach is that it explicitly identifies assumptions about the turbine parameters that guarantee the existence of a solution, but it also helps us to present convergence criteria for different solving algorithms. Glauert's model is also used to optimize blade geometries, in the sense that it maximizes the turbine efficiency. We recall the details of the usual design procedure and briefly discuss the case when a correction of the model is introduced.

Chapter 1

State of the art

This chapter is devoted to an overview of the different subjects discussed in this thesis. We begin with two topics studied in Chapter 2: Data assimilation methods, both in continuous and discrete setting; followed by a summary of space-time parallel methods, in particular the Parareal algorithm. We move on to discuss the key elements of Chapter 3, Bathymetry estimation and Wave modeling, with emphasis on the main assumptions needed to derive different models for wave propagation. It is followed by a brief review of the Blade Element Momentum theory, which is treated in Chapter 4.

1.1 Data assimilation (DA)

Mathematical models are widely used to describe complex systems. They rely on approximations and simplifications of a real phenomenon, which ultimately defines its range of applicability. After proper validation, we would like to apply them to describe or even predict a real situation, as monitoring earthquakes, forecast seasonal influenza or estimate the state of charge of a battery [52]. In order to do so, we need to incorporate real data into our framework. The different techniques that combines mathematical models with available observations to improve the knowledge of a system are known as *Data assimilation* (DA). Among these approaches, here we recall the *Sequential* and *Variational methods*.

1.1.1 Sequential methods

Let us suppose that for some reason, e.g. physical or budgetary constraints, we only have access to partial information of the system. Even in some cases the initial condition is not precisely known, as it happens in climate science [90]. We can treat this lack of information and uncertainty by constructing a new system which uses the available observations to approximate the true state. In a deterministic context, this device is called an *observer*.

We assume that the system is governed by

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \quad t \in [0, +\infty) \\ x(0) = x_0, \\ y(t) = Cx(t), \end{cases} \tag{1.1}$$

where $x \in \mathbb{R}^m$ is the true state vector, $y \in \mathbb{R}^q$ represents the observations (with $q < m$, $m, q \in \mathbb{N}^*$), u is an input and $x(0) = x_0$ an unknown initial condition. The matrices A, B and C are known and their dimensions are consistent.

The *Luenberger observer* [65] imitates the previous model, but includes an extra term in the dynamic that measures the misfit between the observations and its own predictions. It produces a state estimate $\hat{x}$ satisfying

$$\begin{cases} \dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) + L\left[y(t) - \hat{y}(t)\right] & t \in [0, +\infty) \\ \hat{x}(0) = \hat{x}_0, \\ \hat{y}(t) = C\hat{x}(t), \end{cases}$$

with $\hat{x}_0$ an arbitrary initial condition. As long as the original model (1.1) is observable (i.e. the initial state $x(0)$ can be uniquely determined from the observations in $[0, T]$, for any T), the resulting error $\|x(t) - \hat{x}(t)\|$ can be driven to zero at exponential rate by properly choosing the *observer gain* matrix L, and then the true state is recovered asymptotically.

Another alternative is the *Kalman filter* [56], which takes into account measurement errors and model innacuracies represented by Gaussian white noises (both stationary and mutually uncorrelated), in order to construct a state estimate that minimizes the mean square error. Note that extensions to the nonlinear case have been developed, e.g. the *nonlinear Luenberger observer* [3] and *Extended Kalman filter* [53].

A more recent technique, developed by Auroux and Blum [6] is *Back and forth nudging (BFN)*. Nudging simply consists in adding a feedback term in the governing system, as the Luenberger observer does, but also assuming full observations in (1.1) (i.e. $y(t) = Cx_{obs}(t)$, with C invertible) and no input $u(t)$. Using the same ideas, *Backward nudging* considers a bounded time interval $[0, T]$ to approximate instead the initial condition x_0 by solving

$$\begin{cases} \dot{\tilde{x}}(t) = A\tilde{x}(t) - K\left[x_{obs}(t) - \tilde{x}_k(t)\right] & \text{in } [0, T] \\ \tilde{x}(T) = \tilde{x}_T \end{cases}$$

where $\tilde{x}_T$ is an observation of the system at time T and K is the *nudging matrix*, chosen symmetric and positive definite to ensure the asymptotic convergence

$$(\forall t \in (0, T]) \qquad \lim_{\substack{\min \\ \lambda \in \sigma(K)} \lambda \to +\infty} \tilde{x}(t) = x_{obs}(t). \tag{1.2}$$

The BFN algorithm combines these procedures by defining the iterative method

$$\begin{cases} \dot{\hat{x}}_k(t) = A\hat{x}_k(t) + K\left[x_{obs}(t) - \hat{x}_k(t)\right] & \text{in } [0,T] \\ \hat{x}_k(0) = \tilde{x}_{k-1}(0) \end{cases}$$

$$\begin{cases} \dot{\tilde{x}}_k(t) = A\tilde{x}_k(t) - K\left[x_{obs}(t) - \tilde{x}_k(t)\right] & \text{in } [0,T] \\ \tilde{x}_k(T) = \hat{x}_k(T) \end{cases}$$

with $\hat{x}_0(0) = x_0$. Both sequences $\{\hat{x}_k(t)\}_{k\geq 1}$ and $\{\tilde{x}_k(t)\}_{k\geq 1}$ converge to $\hat{x}_\infty(t)$ and $\tilde{x}_\infty(t)$, respectively. Moreover, these limit functions also exhibit the asymptotic behavior described in (1.2), even for $t = 0$.

1.1.2 Variational methods

On the other hand, Sasaki [81] proposed to apply a different approach to meteorology problems: here the governing system constraints a cost functional J, representing the mismatch between the true state $x(t)$ and available data. The goal of this formulation is to minimize $J = J(u)$, where $u(t)$ acts as an input of the governing system (for instance, the initial or boundary condition). In other words, finding the control variable u yields the true state $x = x(u)$. Problems of these kinds fall under the Optimal control theory applied to PDEs, whose theoretical background was originally developed by J.-L. Lions [60].

For illustrative purposes, we consider the continuous minimization problem

$$\begin{aligned} \min_{u\in\mathcal{U}} \quad & J(u) = \frac{1}{2}\int_\Omega (u - x_0^b)^\top B^{-1}(u - x_0^b)dx \\ & \qquad + \frac{1}{2}\int_0^T \int_\Omega (\mathbb{H}(x) - y)^\top R^{-1}(\mathbb{H}(x) - y)\, dxdt \\ \text{s.t.} \quad & \begin{cases} \dot{x}(t) = \mathbb{M}(x(t),t) & \text{in } \Omega \times [0,T] \\ x(0) = u & \text{in } \Omega \end{cases} \end{aligned} \tag{1.3}$$

The initial condition u belongs to a functional space $\mathcal{U}$ that summarizes desirable properties of the control, whereas the definition of $J(u)$ involves several elements: the background and observation error-covariances matrices B and R, which measures the uncertainty around the prior estimate of the initial condition x_0^b and observations $y(t)$, respectively; and a differential operator $\mathbb{H}$ that describes the predicted observations of the governing system. The latter is also modeled by a differential operator $\mathbb{M}$.

We require to compute $\nabla J(u)$, either to derive optimality conditions for (1.3) or to solve numerically this problem. In what follows, we outline different approaches to do so [4, 74, 53].

A direct computation

Since x implicitly depends on u, we must first determine its behavior when the initial condition is slightly perturbed in a direction v. This variation, denoted by $\mathcal{X}$, it satisfies

$$\begin{cases} \dot{\mathcal{X}}(t) = \dfrac{\partial \mathbb{M}}{\partial x}(x(t),t)\mathcal{X} & \text{in } \Omega \times [0,T] \\ \mathcal{X}(0) = v & \text{in } \Omega \end{cases} \tag{1.4}$$

and then, a direct calculation yields

$$\langle \nabla J(u), v \rangle = \int_\Omega \left[B^{-1}(u - x_0^b) + \int_0^T \mathcal{X}^\top \left(\frac{\partial \mathbb{H}}{\partial x}(x(t),t) \right)^* R^{-1}(\mathbb{H}(x) - y)\, dt \right] v\, dx.$$

where * denotes the adjoint operator. The drawback of this method is that requires to solve (1.4) for each direction v.

The Adjoint method

This procedure reduces (1.3) to an unconstrained optimization problem with additional variables, by defining the Lagrangian

$$\mathcal{L}(u,x,\lambda,\mu) = J(u) + \int_0^T \int_\Omega \lambda^\top \left[\dot{x}(t) - \mathbb{M}(x(t),t) \right] \, dxdt + \int_\Omega \mu^\top (x(0) - u)\, dx,$$

where λ and μ are known as the Lagrange multipliers associated with each equation of the governing system.

Instead of dealing with the Tangent-linear equation (1.4), the Lagrangian function allow us to derive an equation for the dual variable of $\mathcal{X}$ in the following way. Setting to zero the derivative of the Lagrangian w.r.t. x in the direction z, at (u,x,λ,μ), leads to

$$\begin{aligned} \langle \nabla_x \mathcal{L}(u,x,\lambda,\mu), z \rangle = 0 \\ \int_0^T \int_\Omega z^\top \left(\frac{\partial \mathbb{H}}{\partial x}(x(t),t) \right)^* R^{-1}(\mathbb{H}(x) - y)\, dxdt \\ \int_0^T \int_\Omega \lambda^\top \left[\dot{z}(t) - \frac{\partial \mathbb{M}}{\partial x}(x(t),t) z \right] \, dxdt + \int_\Omega \mu^\top z(0)\, dx = 0, \end{aligned}$$

and after integrating by parts the second integral and rearranging terms, we obtain

$$\begin{aligned} \int_0^T \int_\Omega z^\top \left[\left(\frac{\partial \mathbb{H}}{\partial x}(x(t),t) \right)^* R^{-1}(\mathbb{H}(x) - y) + \dot{\lambda} - \left(\frac{\partial \mathbb{M}}{\partial x}(x(t),t) \right)^* \lambda \right] dxdt \\ + \int_\Omega \lambda^\top(T) z(T) + \int_\Omega [\mu - \lambda(0)]^\top z(0)\, dx = 0. \end{aligned}$$

Since z is an arbitrary direction, λ must necessarily satisfy the *Adjoint equation*

$$\begin{cases} \dot{\lambda}(t) - \left(\dfrac{\partial \mathbb{M}}{\partial x}(x(t),t)\right)^* \lambda = -\left(\dfrac{\partial \mathbb{H}}{\partial x}(x(t),t)\right)^* R^{-1}(\mathbb{H}(x) - y) & \text{in } \Omega \times [0,T] \\ \lambda(T) = 0 & \text{in } \Omega \end{cases} \tag{1.5}$$

and the supplementary condition $\mu^\top = \lambda(0)$. Then we can compute the gradient for any direction v by

$$\langle \nabla J(u), v \rangle = \langle \nabla_u \mathcal{L}(u, x, \lambda, \mu), v \rangle = \int_\Omega \left[B^{-1}(u - x_0^b) - \lambda^\top(0) \right] v \, dx.$$

Note that we need to solve (1.5) only once, since the adjoint variable does not depend on v.

Discrete variational methods

In contrast to the continuous setting, which requires to solve the tangent-linear equation (1.4) or the adjoint equation to obtain the gradient, discretizing (1.3) leads to its direct computation. Then, discrete variational methods differ in the way they describe the governing system and the treatment of possible nonlinearities.

Given a discretization $\{t_n\}_{n=0}^N$ of $[0,T]$, we denote by x_i and y_i the state and observation vector at time t_i, respectively. Using the same notation as before on the variables J, u, x_0^b, B and R, even if they could depend on the time-discretization considered, the *4D-Var* algorithm [58] reads

$$\begin{aligned} &\min_{u \in \mathcal{U}} J(u) = \frac{1}{2}(u - x_0^b)^\top B^{-1}(u - x_0^b) + \frac{1}{2}\sum_{n=1}^N (\mathcal{H}(x_n) - y_n)^\top R^{-1}(\mathcal{H}(x_n) - y_n) \\ &\text{s.t.} \begin{cases} x_n = \mathcal{M}_{[t_{n-1},t_n]}(x_{n-1}) \quad \forall n = 1, \dots, N \\ x_0 = u \end{cases} \end{aligned} \tag{1.6}$$

where $\mathcal{H}$ is the observation operator and $\{\mathcal{M}_{[t_{n-1},t_n]}(\cdot)\}_{n=1}^N$ is a family of operators that describe the transitions of the state from t_{n-1} to t_n. Then

$$x_n = \left(\mathcal{M}_{[t_{n-1},t_n]} \circ \cdots \circ \mathcal{M}_{[t_0,t_1]}\right)(u) := \mathcal{M}_{[t_0,t_n]}(u)$$

and the gradient can be computed by

$$\nabla J(u) = B^{-1}(u - x_0^b) + \sum_{n=1}^N \boldsymbol{M}_n^\top \boldsymbol{H}_n^\top \cdot R^{-1}(\mathcal{H} \circ \mathcal{M}_{[t_0,t_n]}(u) - y_n). \tag{1.7}$$

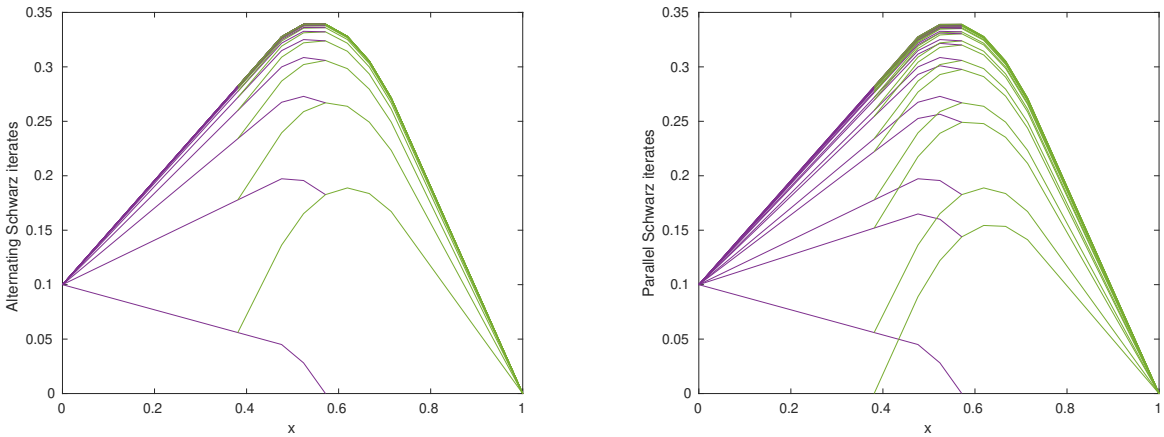
The matrices $\boldsymbol{M}_n = D\mathcal{M}_{[t_0,t_n]}(u)$ and $\boldsymbol{H}_n = D\mathcal{H}(x_n)$ are known as the tangent linear operators associated with $\mathcal{M}_{[t_0,t_n]}$ and $\mathcal{H}$, respectively.

Other variants of this method are *3D-Var*, a time independent version that can be also applied to evolutionary problems by assuming that all the observations are only available at the beginning (i.e. no governing system is needed); the *3D-FGAT*, an improvement of the last in which we replace only $\boldsymbol{M}_n$ by the identity matrix, simplifying the computation of the gradient; and the *Incremental 4D-Var* [24], which consists in approximate (1.6) using a sequence of quadratic minimization problems to reduce the operative cost.

1.2 Space-time parallel methods

For any of the aforementioned DA methods, numerical computation of the state estimate is as relevant as its accuracy. However, the former also requires to be carry out in a reasonable amount of time, which is possible with the help of space or time parallel methods, a natural approach to speed-up the numerical resolution of PDEs using parallel computing. Following Gander [33, 34], we briefly describe some of them.

During the nineteenth century, Fourier analysis was the main tool for studying PDEs, though it is restricted to simple geometries as circles or rectangles. With the purpose of extending Dirichlet's principle to arbitrary domains, Schwarz [82] proposed to solve the Laplace equation by decomposing the domain into two overlapping subdomains where Fourier analysis apply, and then solving alternately a reduced problem on each, in a procedure known nowadays as the *Alternating Schwarz method*. This decomposition is the underlying principle of *Domain decomposition* methods.



(a) Alternating algorithm (b) Parallel algorithm

Figure 1.1: (Discretized) Schwarz methods applied to $-\frac{d^2u}{dx^2}+\eta u = f$ in $[0,1]$ [37, p.8].

A hundred years later, Lions [61, 62, 63] extended the previous method to a parallel framework, by considering possibly overlapping subdomains and solving syncronously the associated local problems. The *Parallel Schwarz method* was born. Since then, in a time when computers become more and more efficient, numerous methods have been developed to take advantage of this strategy. But what about the time dimension? Since the solution of an evolutionary PDE is naturally affected by the past, time is not usually used in parallel computing, even if parallel-in-time methods have been developed for more than 50 years. Its origins can be traced back to Nievergelt [72], who proposed the main idea behind *Multiple shooting* methods: decompose the time interval into disjoint subintervals and solve simultaneously a family of initial-value problems, breaking the intrinsic sequential nature of the time dependent differential equation.

In a more general way, space-time methods differ in the iterative of direct nature of the procedure and the decomposition of the space-time domain considered. Divided into four classes, *Multiple shooting* methods parallelize along the time interval, *Waveform relaxation and Domain decomposition* methods use the space dimension instead and *Space-time multigrid* methods work simultaneously on both, being all of them of iterative nature. In contrast, *Direct time parallel* methods attempt to retrieve a solution by using a direct solver.

In what follows we recall the basics of the Parareal algorithm, one of the most recent Multiple shooting methods; and some examples of parallelization of data assimilation problems.

1.2.1 The Parareal algorithm

The time parallelization of the problem

$$\begin{cases} \dot{u}(t) = f(u(t)), & t \in [0, T] \\ u(0) = u_0 \end{cases}$$

requires to decompose the time interval on N subintervals, denoted by (t_{n-1}, t_n); and introduce intermediate targets that act as initial conditions on each. A direct way to determine these values is solving a nonlinear system of equations by applying Newton's method, an expensive procedure for large systems since it relies on the computation of a Jacobian matrix.

Lions, Maday and Turinici [64] proposed the Parareal algorithm, a different approach which uses two solvers $\mathcal{F}$ and $\mathcal{G}$ that compute a fine and coarse numerical approximation of u on the subintervals and update the artificial initial conditions. Gander and Vandewalle [39] proved that this method reads as a Multiple shooting method in which the Jacobian matrix is approximated by a finite difference in a coarse grid.

The Parareal algorithm approximates $\{u(t_n)\}_{n=1}^N$ by a sequence $\{U_n^k\}_{n=1}^N$, which is contructed as follows:

(a) since the initial conditions are unknown except on the first subinterval, impose arbitrary values on the rest by using the coarse solver $\mathcal{G}$,

$$\begin{aligned} U_n^0 &= \mathcal{G}(t_n, t_{n-1}, U_{n-1}^0), \\ U_0^0 &= u_0. \end{aligned}$$

where $\mathcal{G}(t_n, t_{n-1}, U_{n-1}^0)$ denotes the solution obtained with the coarse at solver t_n, considering U_{n-1}^0 as initial condition at t_{n-1}.

(b) Then solve in parallel the restricted problems

$$\begin{cases} \dot{u}(t) = f(u(t)), \quad t \in [t_{n-1}, t_n] \\ u(t_{n-1}) = U_{n-1}^k \end{cases}$$

using the fine solver $\mathcal{F}$, which yields the approximations $\{\mathcal{F}(t_n, t_{n-1}, U_{n-1}^k)\}_{n=1}^N$.

(c) Finally, smooth the discontinuities previously introduced by defining the sequence

$$U_n^{k+1} := \mathcal{F}(t_n, t_{n-1}, U_{n-1}^k) + \mathcal{G}(t_n, t_{n-1}, U_{n-1}^{k+1}) - \mathcal{G}(t_n, t_{n-1}, U_{n-1}^k),$$

where the superscript k denotes the current number of iterations. Note that on the right-hand side, the first and third term were already computed, whereas the second shows that the update must be done sequentially. Hopefully, this is not computationally expensive since it only depends on the coarse solver $\mathcal{G}$.

The advantage of this algorithm is its superlinear convergence rate. Indeed, due to Gander and Hairer, we have the a priori estimate:

Theorem 1.2.1 (Convergence of Parareal [36]). *Let $\mathcal{F}(t_n, t_{n-1}, U_n^k)$ be the exact solution on the time subdomain (t_{n-1}, t_n) and let $\mathcal{G}(t_n, t_{n-1}, U_n^k)$ be an approximate solution with local truncation error bounded by $C_3 \Delta T^{p+1}$, and satisfying*

$$\mathcal{F}(t_n, t_{n-1}, x) - \mathcal{G}(t_n, t_{n-1}, x) = c_{p+1}(x)\Delta T^{p+1} + c_{p+2}(x)\Delta T^{p+2} + \dots,$$

for ΔT small, where the coefficients $\{c_j\}_{j \geq p+1}$ are continuously differentiable, and assume that $\mathcal{G}$ satisfies the Lipschitz condition

$$\|\mathcal{G}(t + \Delta T, t, x) - \mathcal{G}(t + \Delta T, t, y)\| \leq (1 + C_2 \Delta T) \|x - y\|. \tag{1.8}$$

Then, at iteration k of the Parareal algorithm, we have the bound

$$\left\|u(t_n) - U_n^k\right\| \leq \frac{C_3}{C_1} \frac{(C_1 \Delta T^{p+1})^{k+1}}{k!} (1 + C_2 \Delta T)^{n-(k+1)} \prod_{j=0}^{k} (n - j). \tag{1.9}$$

As a result of the product term in (1.9), after k iterations of the Parareal algorithm the approximation is exact on the first k subintervals (as shown in Figure 1.2), and hence it converges in at most N iterations.

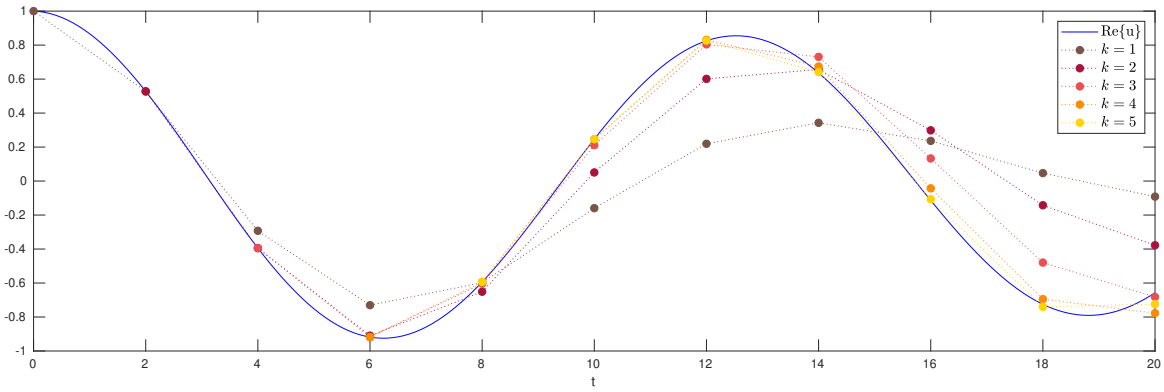


Figure 1.2: Parareal algorithm applied to the Dahlquist equation $\dot{u}(t) = -\frac{\mathrm{i}}{2}u$ in $[0, 20]$

Even though it is not explicitly stated, the constant C_2 must be positive, which is why the Lipschitz assumption (1.8) does not take into account the decaying case, i.e. when the coarse solver is contracting. Since we are interested in coupling this algorithm with the Luenberger observer and take advantage of its decaying behavior, in Chapter 2 we present a variant of Theorem 1.2.1 that covers this situation.

1.2.2 Space-time parallel methods and DA

Trémolet and Le Dimet [88] were among the first to address the parallelization of Variational data assimilation problems in meteorology. In a continuous setting, they proposed a Domain decomposition approach combined with the Adjoint method, by assigning to each subdomain a local version of (1.3), with an extra term on the local cost functional to enforce the continuity of the state between adjacent domains.

Twenty years later, new strategies in this topic also include the time direction. For instance, Rao and Sandu [79] apply a quasi-Newton solver to the 4D-Var problem, but they time-parallelize first the computation of the gradient. A more sophisticated approach is proposed by D'Amore and Cacciapuoti [26], who combines the Parareal algorithm with the Multiplicative Parallel Schwarz method (MPS) to solve 4D-Var.

The methodology associated with the Parareal algorithm has also been applied to optimization problems, but not necessarily related to DA. For instance, Maday, Salomon and Turinici [66] proposed a specific algorithm to solve optimality systems in the case of quantum control. By defining intermediate states and then solving a family of local optimization problems in parallel, these values are updated after solving sequentially the forward and adjoint equations, which is computationally cheap compared with the optimization procedure. Since the original cost functional is not parallelizable, the method uses a different one which depends on the adjoint variable and can be also decomposed as the sum of cost functionals for each subinterval. Ultimately, its optimal solution coincides with that of the original problem.

1.3 Bathymetry estimation

Recently, considerable literature has grown up around the subject of inverse problems in free surface flows. A review from Sellier identifies different techniques applied for bathymetry reconstruction [83, Section 4.2], which rely mostly on the derivation of an explicit formula for the bathymetry, numerical resolution of a governing system, or variational data assimilation [51].

A variational approach is also well suited for solving coastal engineering problems, since they must satisfy several mechanical constraints. For instance, among the several aspects to consider when designing a harbor, building defense structures is essential to protect it against wave impact. These structures can be optimized to minimize the wave energy, by studying its interaction with the reflected waves [55]. Bouharguane and Mohammadi [69, 14] consider a time-dependent approach to study the evolution of sand motion at the seabed, which could also allow these structures to change in time. In this case, the proposed functionals are locally minimized using sensitivity analysis, a technique broadly applied in geosciences.

From a mathematical point of view, the solving of these kinds of problem is mostly numerical. Questions about a suitable control space for the bathymetry, controllability, regularity or well-posedness of the solution are not usually adressed. A theoretical approach applied to the modeling of surfing pools can be found in [25, 71], where the goal is to maximize the total energy of the prescribed wave. The former proposes to determine a bathymetry, whereas the latter sets the shape and displacement of an underwater object along a constant depth.

1.3.1 Wave modeling

In all the preceding problems, a crucial constraint for determining the bathymetry is its interaction with sea waves. The cornerstone of its modeling is the Navier-Stokes system

$$\begin{cases} \rho\left(\dfrac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\,\mathbf{u}\right) = \operatorname{div}(\sigma_T) + \mathbf{g} & \text{in } \Omega_t, \\ \operatorname{div}(\mathbf{u}) = 0 & \text{in } \Omega_t, \\ \mathbf{u} = \mathbf{u}_0 & \text{in } \Omega_0, \end{cases}$$

where $\mathbf{u} = (u, v, w)^\top$ denotes the velocity of the fluid in the bounded and time-dependent region Ω_t; ρ, $\mathbf{g}$ and σ_T represent the density coefficient, gravity and total stress tensor, respectively. Here we omit the boundary conditions on the water level $\eta(x,t)$ and at the bottom $-z_b(x)$, possibly depending on the pressure, viscosity and friction effects, see Section 3.1.1 for more details. In what follows, for simplicity we neglect the last two.

Certain simplifications of the Navier-Stokes system gives rise to different models for wave propagation [16, 59], all of them depending on the relationship between three characteristic parameters known as the relative depth H, the horizontal lenght L and the maximum vertical amplitude A. To recall some of them, we introduce the ratios

$$\varepsilon = \frac{H}{L},\ \delta = \frac{A}{H}.$$

Vertically averaged fluids as rivers, coastal domains and oceans can be modeled by using the shallow water assumption $\varepsilon \ll 1$, whereas small amplitude waves are described by models where $\delta \ll 1$.

Assuming a shallow water regime, as well as different values of δ and approximations of the pressure, leads to the following models.

Saint-Venant system

The *hydrostatic pressure* approximation (i.e. the contribution of the vertical acceleration of the fluid in the pressure is negligible) combined with $\delta = \mathcal{O}(1)$ yields

$$\begin{cases} \dfrac{\partial \eta}{\partial t} + \dfrac{\partial h\overline{u}}{\partial x} + \dfrac{\partial h\overline{v}}{\partial y} = 0 \\ \dfrac{\partial h\overline{u}}{\partial t} + \dfrac{\partial h\overline{u}^2}{\partial x} + \dfrac{\partial h\overline{u}\,\overline{v}}{\partial y} + gh\dfrac{\partial \eta}{\partial x} = 0 \\ \dfrac{\partial h\overline{v}}{\partial t} + \dfrac{\partial h\overline{u}\,\overline{v}}{\partial x} + \dfrac{\partial h\overline{v}^2}{\partial x} + gh\dfrac{\partial \eta}{\partial y} = 0 \end{cases} \tag{1.10}$$

where $h = \eta + z_b$ is the water height and $\overline{u}$, $\overline{v}$ are depth-averaged velocities. This model is commonly used for modeling dam breaks [40], hydraulic jumps and tidal flows in estuaries.

Wave equation

Keeping the hydrostatic pressure but assuming instead $\delta \ll 1$ is equivalent to neglect the convective acceleration terms in (1.10) and linearize the new system around the sea level $\eta = 0$. Combining the resulting expressions brings

$$\frac{\partial^2 \eta}{\partial t^2} - \nabla\left(g z_b \nabla \eta\right) = 0.$$

In particular, its solution has the form $\eta(x,t) = \mathrm{Re}\{\psi_{tot}(x)e^{-i\omega t}\}$, where ω is a given frequency and the amplitude ψ_{tot} satisfies the *Helmholtz equation*

$$\triangle \psi_{tot} + \left(\frac{\omega^2}{g z_b}\right)\psi_{tot} = 0,$$

when constant depth $-z_b$ is assumed. Under suitable boundary conditions, this equation is used for studying water-wave scattering problems, such as the one presented in Chapter 3, where a variable bathymetry acts as a scatterer.

Classical Boussinesq system

On the other hand, a non-hydrostatic pressure introduces dispersive terms in the governing equations. A third ratio, called the Ursell number

$$U_r = \frac{\delta}{\varepsilon^2},$$

measures their strength with respect to the nonlinear terms [89]. In two-dimensional motion, a specific choice of this parameter ($U_r = \mathcal{O}(1)$ and $\delta \ll 1$) leads to

$$\begin{cases} \dfrac{\partial \eta}{\partial t} + \dfrac{\partial h\overline{u}}{\partial x} = 0 \\ \dfrac{\partial \overline{u}}{\partial t} + \overline{u}\dfrac{\partial \overline{u}}{\partial x} + g\dfrac{\partial \eta}{\partial x} = \dfrac{z_b^2}{3}\dfrac{\partial^3 \overline{u}}{\partial x^2 \partial t} \end{cases}$$

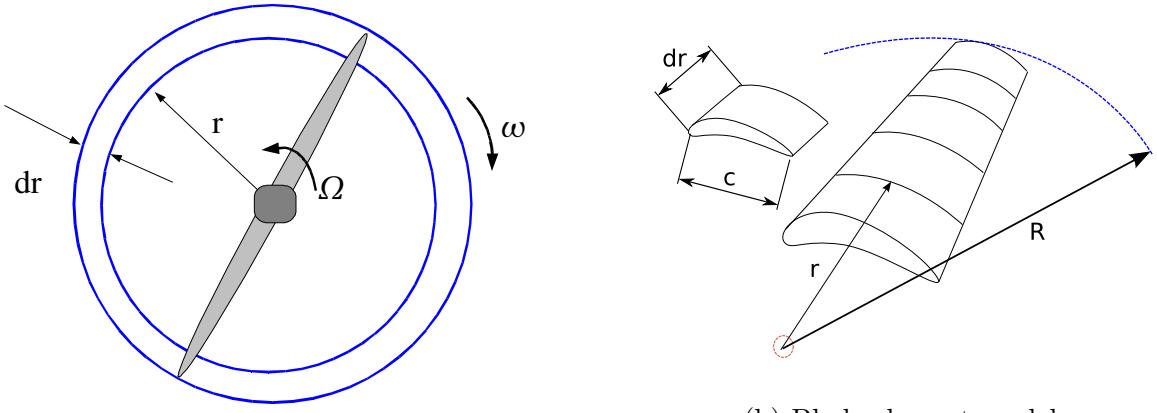
for a constant depth $-z_b$. This model and its subsequent variants are used to describe buoyant motion in fluids, as waves entering the near-shore zone [10] or landslide generated waves [28].

1.4 Blade element momentum (BEM) theory

Blade element momentum (BEM) theory is a mechanical model widely used to evaluate turbine performance, according to the mechanical/geometric characteristics of their blades and the current to which they are exposed. Developed by Glauert [44], it follows from the combination of two different models: *Blade element theory* (BET) and *Momentum theory* (MT).

By cutting the blade into sections that are treated according to a planar model, Blade element theory studies the turbine behavior from a local point of view [32]. The fundamental quantities of this model are the coefficients C_L and C_D called *drag* and *lift*, which are introduced to account for the drag and lift forces expressed in the cut plane. The results are then integrated along the blade to obtain the quantities of global interest. In constrast, Momentum theory [78] (also known as disk actuator theory or Axial momentum theory) is a global theory that macroscopically studies the behavior of a fluid column passing through a turbine.

BEM theory then relies on a decomposition of the fluid/turbine system into a macroscopic part via the MT and a local planar part via the BET. The latter considers a radial decomposition for the blades and fluid column (Figure 1.3a), by splitting the rotor area into concentric rings of infinitesimal thickness that do not interact with each other.



(a) Radial decomposition

(b) Blade element model

Figure 1.3: Decompositions involved in BEM Theory [54, p.8].

Glauert's model ultimately links three variables associated with the ring under consideration: the *axial induction factor* a, the *tangential induction factor* a' and the *relative angle deviation* φ. Given a blade profile and assuming a fixed rotation speed, the resulting equations are

$$\tan\varphi = \frac{1-a}{\lambda(1+a)}, \tag{1.11}$$

$$\frac{a}{1-a} = \frac{1}{4\sin^2\varphi}\cdot\frac{Bc_\lambda}{2\pi r}\left(C_L(\varphi-\gamma_\lambda)\cos\varphi + C_D(\varphi-\gamma_\lambda)\sin\varphi\right), \tag{1.12}$$

$$\frac{a'}{1-a} = \frac{1}{4\lambda\sin^2\varphi}\cdot\frac{Bc_\lambda}{2\pi r}\left(C_L(\varphi-\gamma_\lambda)\sin\varphi - C_D(\varphi-\gamma_\lambda)\cos\varphi\right). \tag{1.13}$$

where r denotes the distance between the blade element taken into account (Figure 1.3b) and the rotor, $\lambda = \lambda(r)$ is the tip-speed ratio and B the total number of blades. The *chord* c_λ and *twist angle* γ_λ are given by the blade geometry.

Various modifications have been introduced to the model to consider e.g. the flow around the tip of a blade or a turbulent wake state. For a more detailed account, see Section 4.1.4.

Finally, we evaluate the turbine efficiency using the *power coefficient*

$$C_p(\varphi) = \frac{8}{\lambda_{\max}}\int_{\lambda_{\min}}^{\lambda_{\max}} \lambda^3 a'(1-a)\left(1-\frac{C_D}{C_L}(\varphi-\gamma_\lambda)\tan^{-1}\varphi\right)d\lambda.$$

Conversely, we can design a blade profile that maximizes this quantity by solving

$$\begin{aligned} \max_{\varphi}\ & C_p(\varphi) \\ s.t\ & (1.11\text{-}1.13) \\ & \overline{\alpha} = \varphi - \gamma_\lambda, \end{aligned}$$

where $\overline{\alpha}$ is chosen such that the ratio $\frac{C_D}{C_L}$ is close to zero. In order to simplify this optimization problem, it is commonly assumed a zero drag coefficient, which leads to an explicit solution. To the best of our knowledge, the general case concerning non-zero drag and possible corrections of the model has not been addressed.

Chapter 2

Time-parallelization of sequential data assimilation problems

We present in this chapter a first algorithm to parallelize in time a Luenberger observer. By dividing the time interval into a sequence of data processing *windows*, we then perform the time-parallelization on each, following a stopping criterion that preserves the exponential rate of convergence of the observer. In the particular case of the Parareal algorithm, we obtain an a priori estimate of the number of parareal iterations required on each window. The efficiency of the procedure is also studied. We present experiments to confirm the results obtained theoretically.

This is a joint work with Felix Kwok (Hong Kong Baptist University) and Julien Salomon (ANGE, INRIA Paris).

2.1 The Luenberger observer

Control theory usually requires total knowledge of the state vector. However, due to certain limitations related to a problem, for instance the number of available measurements, one can often have access only to partial information. A dynamic which fits into this setting is given by

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0 \\ y(t) = Cx(t) \end{cases} \tag{2.1}$$

where $A \in \mathcal{M}_{m\times m}(\mathbb{R})$, $B \in \mathcal{M}_{m\times p}(\mathbb{R})$ and $C \in \mathcal{M}_{q\times m}(\mathbb{R})$ are assumed to be known (with $m, p, q \in \mathbb{N}^*$, $p, q < m$); $x \in \mathbb{R}^m$ is the state vector, $y \in \mathbb{R}^q$ is the measured output, $u \in \mathbb{R}^p$ and $t \in (0, +\infty)$. The initial condition $x(0) = x_0$ is unknown.

We are interested in computing an estimate $\hat{x}(t)$ of $x(t)$, knowing only the imput $u(t)$ and output $y(t)$. To tackle this problem, Luenberger [65, pp.300-307] proposed the model

$$\begin{cases} \dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) + L\left[y(t) - \hat{y}(t)\right], \quad \hat{x}(0) = \hat{x}_0 \\ \hat{y}(t) = C\hat{x}(t) \end{cases} \tag{2.2}$$

with $L \in \mathcal{M}_{m\times q}(\mathbb{R})$ the *observer gain*. System (2.2) is known as the *Luenberger observer* or the *Identity observer*.

The matrix L needs to be specified, but let us already note that it plays an important role in the estimation error $x(t) - \hat{x}(t)$. Indeed, substracting (2.1) and (2.2), and then solving the resultant ODE, we get

$$x(t) - \hat{x}(t) = \mathrm{e}^{(A-LC)t}\,(x(0) - \hat{x}(0)) \tag{2.3}$$

If the eigenvalues of $A - LC$ lie in the open left half-plane $\{z \in \mathbb{C} : \mathrm{Re}\{z\} < 0\}$, then the error will decay to zero. Due to the following result, known as the *Identity observer Theorem* [65, p.303], we can construct a matrix L such that this property holds:

Theorem 2.1.1. *Given a completely observable system (2.1), an identity observer of the form (2.2) can be constructed, and the coefficients of the characteristic polynomial of the observer can be selected arbitrarily.*

To be more precise, let us recall that System (2.1) is *observable* if the rank of the matrix

$$\mathcal{C} := \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{m-1} \end{bmatrix}$$

is m. Then, given a set $(\mu_i)_{i=1,\dots,m}$, Theorem 2.1.1 guarantees the existence of L that satisfies

$$\det\left(sI - (A - LC)\right) = \phi(s) \tag{2.4}$$

where $\phi(s) = (s - \mu_1)\cdots(s - \mu_m)$, i.e. $(\mu_i)_{i=1,\dots,m}$ are the eigenvalues of $A - LC$.

Note that for a single-input single-output system, i.e. $p = q = 1$, we could determine a unique $L \in \mathbb{R}^m$ by equating the m coefficients of both polynomials in (2.4). However, this approach leads to highly nonlinear equations that are in practice not tractable. Another way to proceed is the Bass-Gura method [12], which requires the first companion form of A and the coefficients of $\phi(s)$. An even more direct method consists in using the Ackermann formula [1] for an observable system, to define L by

$$L = \phi(A)\mathcal{C}^{-1}(0 \cdots 0 \;\; 1)^\top$$

which is a consequence of the Cayley-Hamilton Theorem. For its multi-input multi-output extension, see [2].

Due to Theorem 2.1.1, we obtain

Proposition 2.1.2. *We assume System (2.1) is observable and the eigenvalues of $A - LC$ are negative and simple. Then, we have*

$$\left\|\mathrm{e}^{(A-LC)t}\right\| \leq \gamma \mathrm{e}^{-\mu t}$$

with $\mu := \min_{\nu \in \sigma(A-LC)} |\nu|$ and $\gamma := cond(V) = \|V^{-1}\|\,\|V\|$, where V is the matrix whose rows are the eigenvectors of $A-LC$ and $\|\cdot\|$ represents the induced 2-norm of a matrix.

In particular, combining the latter with Equation (2.3) yields

$$\|x(t) - \hat{x}(t)\| \leq \gamma \, \|x(0) - \hat{x}(0)\| \, \mathrm{e}^{-\mu t}. \tag{2.5}$$

In practice, the term $\|x(0) - \hat{x}(0)\|$ is unknown, whereas μ is constructed (and known) by the procedure that designs L. Consequently, this equation only provides a rate of convergence for the Luenberger observer.

2.2 Time-parallelization setting

In the last decades many techniques have been designed to combine space-time parallelization procedures with data assimilation or optimal control algorithms on bounded time intervals, for instance [88, 79, 26, 66]. In what follows, we propose to extend this notion to the unbounded case and at the same time exploit the exponential rate of convergence of the problem, by an approach that we call the *Diamond strategy*.

Let us describe briefly our approach. We proceed by partitioning $(0, +\infty)$ on intervals of the same length that we call *windows*. Following a sequential order, we apply a parallel-in-time solver on each, up to some level of accuracy related to a specific stopping criterion. Since our analysis allows us to decompose the estimation error into two terms, corresponding respectively to the Luenberger observer and the parallelization error, we propose a suitable bound on the latter, so that our stopping criterion preserves Luenberger's rate of convergence.

2.2.1 Framework

In order to accelerate the assimilation and take advantage of a time-parallelization procedure, we propose to divide the time interval into windows of a given length $T > 0$, denoted by

$$W_\ell := (T_\ell, T_{\ell+1}), \quad \ell \in \mathbb{N}$$

where $T_\ell = \ell \cdot T$. Then, we solve Equation (2.2) on each window, in a sequential order, using a time-parallel algorithm.

Let us describe how the latter is applied. Given $\ell \in \mathbb{N}$ and a fixed window W_ℓ, we decompose it into N subintervals of length ΔT

$$W_\ell = \bigcup_{n=0}^{N-1} (t_n^\ell, t_{n+1}^\ell)$$

with $t_n^\ell = T_\ell + n\Delta T$ and $N\Delta T = T$. Since time moves forward, parallelizing in this direction requires the introduction of initial conditions $\hat{X}_{\ell,n}^h$ on each subinterval, which are obtained from the time-parallelization procedure under consideration. In this setting, the parameter h is related with the accuracy of the method.

When $n = 0$, we consider as initial condition $\hat{X}^h_{0,0} = \hat{x}_0$ in the first window, and $\hat{X}^h_{\ell,0} = \hat{x}_\parallel(T_\ell^-)$ for $\ell > 0$. Finally, we construct a parallel version of Equation (2.2) in each subinterval (t_n^ℓ, t_{n+1}^ℓ) through the equation

$$\begin{cases} \dot{\hat{x}}_\parallel(t) = A\hat{x}_\parallel(t) + Bu(t) + L\left[y(t) - C\hat{x}_\parallel(t)\right] \\ \hat{x}_\parallel(t_n^{\ell\,+}) = \hat{X}^h_{\ell,n} \end{cases} \tag{2.6}$$

where $\hat{x}_\parallel(t)$ denotes the approximation of $\hat{x}(t)$ obtained by the parallel-in-time solver under consideration.

2.2.2 The Diamond strategy

Imposing initial conditions induces discontinuities at the interfaces t_n^ℓ of the sub-intervals. In what follows, we call these differences *jumps*, defined by $J^h_{\ell,n} := \hat{X}^h_{\ell,n} - \hat{x}_\parallel(t_n^{\ell\,-})$. The success of the parallel method relies on their decay to zero. Note that jumps can be computed without knowing the true solution and hence be used to design an a posteriori estimate of the method.

In this setting, we can express the relation between the solution of (2.1) and its parallel version (2.6) by

Lemma 2.2.1. *Under the assumptions of Proposition 2.1.2, we have*

$$\left\|\varepsilon_\parallel(t_n^{\ell\,-})\right\| \leq \gamma\left(\mathrm{e}^{-\mu t_n^\ell}\left\|x(0) - \hat{x}(0)\right\| + \mathrm{e}^{-\mu\Delta T}\left\|J^h_{\ell,n-1}\right\|\right)$$

where $\varepsilon_\parallel(t) = x(t) - \hat{x}_\parallel(t)$ is the error of approximation (2.6) and $t_n^\ell = (N\ell + n)\Delta T$.

Proof. Since this result relies only on the subintervals and does not depend on the windows, we simplify the notation using the sequence $(t_i)_{i\in\mathbb{N}}$, with $i = N\ell + n$.

Let $i \in \mathbb{N}^*$. Substracting Equation (2.6) and (2.1), the latter restricted to the sub-interval (t_i, t_{i+1}), we get

$$\begin{cases} \dot{\varepsilon}_\parallel(t) = (A - LC)\varepsilon(t) \\ \varepsilon_\parallel(t_i^+) = \hat{X}^h_{\ell,i} - x(t_i) \end{cases}$$

Its solution is given by

$$\varepsilon_\parallel(t) = \mathrm{e}^{(A-LC)(t-t_i)}\varepsilon_\parallel(t_i^+).$$

Setting $t = t_{i+1}^-$ in the equation above, we get

$$\varepsilon_\parallel(t_{i+1}^-) = \mathrm{e}^{(A-LC)\Delta T}(\varepsilon_\parallel(t_i^+) - J_i^h) + \mathrm{e}^{(A-LC)\Delta T}J_i^h \tag{2.7}$$

On the other hand, from the definition of jump, we have

$$\begin{aligned}\varepsilon_{\|}(t_i^+) - J_i^h &= \hat{x}_{\|}(t_i^-) - x(t_i) = \mathrm{e}^{(A-LC)\Delta T}\mathrm{e}^{-(A-LC)\Delta T}(\hat{x}_{\|}(t_i^-) - x(t_i)) \\ &= \mathrm{e}^{(A-LC)\Delta T}(\hat{x}_{\|}(t_{i-1}^-) - x(t_{i-1})) = \mathrm{e}^{(A-LC)\Delta T}(\varepsilon_{\|}(t_{i-1}^+) - J_{i-1}^h).\end{aligned}$$

Applying the previous formula recursively in (2.18) leads to

$$\varepsilon_{\|}(t_{i+1}^-) = \mathrm{e}^{(A-LC)(i+1)\Delta T}\varepsilon_{\|}(0^+) + \mathrm{e}^{(A-LC)\Delta T}J_i^h,$$

since $J_0^h = 0$. The result follows from taking the norm and using Proposition 2.1.2. □

We recall now that our approach aims to preserve Luenberger's rate of convergence. Thanks to the previous lemma, we can achieve this goal by defining on each window W_ℓ the stopping criterion

$$\max_{1\le n\le N}\left\|J_{\ell,n}^h\right\| \le \widetilde{\gamma}\mathrm{e}^{-\mu\ell T},$$

see Proposition 2.2.2 (on next page) for more details. We summarize our procedure as follows.

Algorithm 2.1: Diamond strategy

Input: $A, C, \hat{x}_0, T, N, (\mu_i)_{i=1,\dots m}, \widetilde{\gamma}$

Output: $(t_n^\ell)_{\substack{n=0,\dots,N\\ \ell\in\mathbb{N}}}, (\hat{X}_{n,\ell}^h)_{\substack{n=0,\dots,N\\ \ell\in\mathbb{N}}}$

`/* place denotes a function to construct L, as in Theorem 2.1.1 */`

$L = \texttt{place}(A, C, (\mu_i)_{i=1,\dots,m})$, $\mu = \min_{\nu\in\sigma(A-LC)}|\nu|$

$\Delta t := \frac{T}{N}$

$\ell = 0$

repeat

  $T_\ell = \ell T$

  $(\forall n \in \{1,\dots,N\})\quad t_n^\ell = T_\ell + n\Delta T$

  if $\ell = 0$ **then**

    $\hat{X}_{\ell,0}^h = \hat{x}_0$

  else

    $\hat{X}_{\ell,0}^h = \hat{x}_{\|}(T_\ell^-)$

  end

  `/* Using a generic time-parallelization procedure (GTP): */`

  Construct the initial conditions $\{\hat{X}_{\ell,n}^h\}_{n=0,\dots,N-1} = \texttt{GTP}(\hat{X}_{\ell,0}^h)$

  $J_{\ell,n}^h := \hat{X}_{\ell,n}^h - \hat{x}_{\|}(t_n^{\ell\,-})$

  Assign $\ell \leftarrow \ell + 1$

until $\max_{1\le n\le N}\left\|J_{\ell,n}^h\right\| \le \widetilde{\gamma}\mathrm{e}^{-\mu\ell T}$

Proposition 2.2.2. *Let us assume that h is obtained from the stopping criterion in W_ℓ*

$$\max_{1\leq n\leq N} \left\| J^h_{\ell,n} \right\| \leq \widetilde{\gamma} \mathrm{e}^{-\mu \ell T} \tag{2.8}$$

where $\widetilde{\gamma}$ is an arbitrary parameter. Then, the rate of convergence of $\hat{x}_\parallel(t)$ to $x(t)$ is bounded by μ, i.e.

$$\left\| \varepsilon_\parallel(t_n^{\ell\,-}) \right\| \leq \gamma \mathrm{e}^{-\mu \Delta T} \left(\|x(0) - \hat{x}(0)\| + \widetilde{\gamma} \right) \mathrm{e}^{-\mu \ell T}.$$

2.3 Parallelization

Independent to the choice of the parallel-in-time solver, Algorithm 2.1 is well-defined since the jumps are computed a posteriori. However, by specifying it, we can study in more detail the stopping criterion presented in (2.8) and the complexity of the overall procedure. In order to obtain an a priori estimate for the required accuracy h on each window and the efficiency of the *Diamond strategy*, throughout this section we consider the Parareal algorithm as the time-parallel method.

2.3.1 The Parareal algorithm

Introduced by Lions, Maday and Turinici [64], the goal of the Parareal algorithm is to solve evolution problems by partitioning a bounded time interval into subintervals, and after assigning to each of them a processor, updates iteratively the initial conditions and solve a series of independent and smaller problems in parallel, reducing the time-computation of the solution.

To be more precise, given the problem

$$\begin{cases} \dot{u}(t) = f(u(t)), \quad t \in [0,T] \\ u(0) = u_0 \end{cases} \tag{2.9}$$

we divide $[0,T]$ in M subintervals denoted by (t_{n-1}, t_n). Then we consider two solvers $\mathcal{F}$ and $\mathcal{G}$, that compute a fine and a coarse numerical approximation of u. The former is computationally expensive and consequently restricted to solve initial-value problems with high accuracy in each subinterval (t_{n-1}, t_n), whereas the latter is faster and can be used for solving on large intervals as $[0,T]$. For an arbitrary initial condition $\tilde{u}$ given in $t = t_{n-1}$, we denote these local approximations of $u(t_n)$ by $\mathcal{F}(t_n, t_{n-1}, \tilde{u})$ and $\mathcal{G}(t_n, t_{n-1}, \tilde{u})$, respectively.

In this framework, $(u(t_n))_{n=1,\dots,M}$ is approximated by $(U_n^k)_{n=1,\dots,M}$, which is build as follows:

Algorithm 2.2: Parareal algorithm

Input: u_0, T, M, Tol
Output: $(t_n)_{n=1,\dots,M}, (U_n^{k^*})_{n=1,\dots,M}$
$\Delta t := \frac{T}{N}, t_0 = 0$
`/* Initialization of the initial conditions */`
$U_0^0 = u_0$
for $1 \leq n \leq M$ **do**
  $t_n = n\Delta T$
  $U_n^0 = \mathcal{G}(t_n, t_{n-1}, U_{n-1}^0)$
end

$k = 0$
repeat
  $U_0^k = u_0$
  for $1 \leq n \leq M$ **do**

$$U_n^{k+1} = \mathcal{F}(t_n, t_{n-1}, U_{n-1}^k) + \mathcal{G}(t_n, t_{n-1}, U_{n-1}^{k+1}) - \mathcal{G}(t_n, t_{n-1}, U_{n-1}^k) \quad (2.10)$$

  end

  $J_n^k := U_n^k - u(t_n^-)$
  Assign $k \leftarrow k + 1$
until $\max_{1 \leq n \leq M} \left\| J_n^k \right\| \leq \text{Tol}$
$k^* = k - 1$

Notice that the superscript k in Algorithm 2.2 plays the role of the parameter h, introduced in the previous section.

As a remark, Gander and Vandewalle [39] showed that the parareal algorithm can be presented as a multi-shooting algorithm applied to (2.9). Solving the multiple shooting equations with the Newton's method yields

$$U_n^{k+1} = u_{n-1}(t_n, U_{n-1}^k) + \frac{\partial u_{n-1}}{\partial U_{n-1}}(t_n, U_{n-1}^k)(U_{n-1}^{k+1} - U_{n-1}^k),$$

where $u_{n-1}(t_n, U_{n-1}^k)$ denotes the exact solution of (2.9) at t_n, with initial condition U_{n-1}^k at t_{n-1}. If we approximate the exact solution using the fine solver and the Jacobian term by a difference on a coarse grid, we obtain Equation (2.10).

Concerning its convergence, they proved that it requires at most M iterations. Moreover, after k iterations the algorithm is exact on the first k subintervals. An improvement of their result, due to Gander and Hairer [36], assumes that the coarse solver must satisfy

$$\|\mathcal{G}(t_n, t_{n-1}, y) - \mathcal{G}(t_n, t_{n-1}, z)\| \leq (1 + C_2\Delta T)\|y - z\|,$$

for a positive constant C_2. However, this result does not cover the decaying case, i.e. when the Lipschitz constant is smaller than one. Since we are interested in coupling this algorithm with the Luenberger observer and take advantage of its decaying behavior, we provide a result adapted to this case, which follows from [38].

Theorem 2.3.1 (Convergence of Parareal for decaying problems)**.** *Let $\mathcal{F}(t_n, t_{n-1}, U_{n-1}^k)$ be the exact solution on the time subdomain (t_{n-1}, t_n) and let $\mathcal{G}(t_n, t_{n-1}, U_{n-1}^k)$ be a coarse integrator, such that the local truncation error $\tau(t_n, z) := \mathcal{F}(t_n, t_{n-1}, z) - \mathcal{G}(t_n, t_{n-1}, z)$ satisfies for all z*

$$\begin{aligned}\|\tau(t_n, z)\| &\leq \alpha,\\ \|\tau(t_n, y) - \tau(t_n, z)\| &\leq \beta\|y - z\|,\end{aligned}$$

where $\alpha, \beta > 0$ are constants. We also assume that $\mathcal{F}$ and $\mathcal{G}$ are Lipschitz with respect to the initial conditions:

$$\max\left\{\|\mathcal{F}(t_n, t_{n-1}, y) - \mathcal{F}(t_n, t_{n-1}, z)\|, \|\mathcal{G}(t_n, t_{n-1}, y) - \mathcal{G}(t_n, t_{n-1}, z)\|\right\} \leq \varepsilon\|y - z\|,$$

for a constant $\varepsilon \in (0, 1)$.

Then, after k iterations of the Parareal algorithm, the error $\left\|U_n^k - u(t_n)\right\|$ is bounded above by E_n^k, defined as

$$E_n^k = \begin{cases} 0 & n \leq k \\ \alpha\beta^k \sum_{i=0}^{n-k-1} \binom{k+i}{k} \varepsilon^i & n > k. \end{cases} \tag{2.11}$$

Proof. From the definition of the Parareal algorithm in (2.10) and since $\mathcal{F}$ is the exact solution in (t_{n-1}, t_n), we obtain, after adding and substracting $\mathcal{G}(t_n, t_{n-1}, u(t_{n-1}))$

$$\begin{aligned} U_n^k - u(t_n) &= \mathcal{F}(t_n, t_{n-1}, U_{n-1}^{k-1}) + \mathcal{G}(t_n, t_{n-1}, U_{n-1}^k) - \mathcal{G}(t_n, t_{n-1}, U_{n-1}^{k-1}) \\ &\quad - \mathcal{F}(t_n, t_{n-1}, u(t_{n-1})) \\ &= \tau(t_n, U_{n-1}^{k-1}) - \tau(t_n, u(t_{n-1})) + \mathcal{G}(t_n, t_{n-1}, U_{n-1}^k) - \mathcal{G}(t_n, t_{n-1}, u(t_{n-1})), \end{aligned}$$

and then, taking norms and using the first assumption on τ, combined with the Lipschitz condition on $\mathcal{G}$, leads to

$$\left\|U_n^k - u(t_n)\right\| \leq \beta\left\|U_{n-1}^{k-1} - u(t_{n-1})\right\| + \varepsilon\left\|U_{n-1}^k - u(t_{n-1})\right\|.$$

Moreover, the error in the initial guess can be estimated in the same way, although using the coarse integration $U_n^0 = \mathcal{G}(t_n, t_{n-1}, U_{n-1}^0)$, which gives

$$\left\|U_n^0 - u(t_n)\right\| \leq \alpha + \varepsilon \left\|U_{n-1}^0 - u(t_{n-1})\right\|.$$

Therefore, an upper bound E_n^k for $\left\|U_n^k - u(t_n)\right\|$ satisfies the recurrence relation

$$E_n^k = \beta E_{n-1}^{k-1} + \varepsilon E_{n-1}^k, \tag{2.12}$$
$$E_n^0 = \alpha + \varepsilon E_{n-1}^0, \tag{2.13}$$

with $E_0^k = 0$ for all k. We solve this recurrence using a generating function, by defining the formal power series

$$\rho_k(\zeta) = \sum_{n\geq 1} E_n^k \zeta^n.$$

Multiplying (2.12) and (2.13) by ζ^n and summing over $n \geq 1$ gives

$$\rho_k(\zeta) = \beta\zeta\rho_{k-1}(\zeta) + \varepsilon\zeta\rho_k(\zeta),$$
$$\rho_0(\zeta) = \frac{\alpha\zeta}{1-\zeta} + \varepsilon\zeta\rho_0(\zeta),$$

and then, solving by induction yields the explicit formula

$$\rho_k(\zeta) = \frac{\alpha\beta^k\zeta^{k+1}}{(1-\zeta)(1-\varepsilon\zeta)^{k+1}}.$$

Expanding $\rho_k(\zeta)$ in a power series leads to

$$\rho_k(\zeta) = \alpha\beta^k\zeta^{k+1}\Big(\sum_{i\geq 0}\zeta^i\Big)\Big(\sum_{j\geq 0}\binom{k+j}{k}(\varepsilon\zeta)^j\Big) = \alpha\beta^k\zeta^{k+1}\sum_{n\geq 0}\Big(\sum_{i=0}^{n}\binom{k+i}{k}\varepsilon^i\Big)\zeta^n$$
$$= \sum_{n\geq 0}\Big(\alpha\beta^k\sum_{i=0}^{n}\binom{k+i}{k}\varepsilon^i\Big)\zeta^{n+k+1}.$$

Then, for $n \leq k$ we have $E_0^k = \ldots = E_k^k = 0$; whereas for $n > k$, we obtain

$$E_n^k = \alpha\beta^k\sum_{i=0}^{n-k-1}\binom{k+i}{k}\varepsilon^i. \qquad \square$$

2.3.2 Combination with Luenberger observer

For a fixed window W_ℓ, we approximate $\hat{x}_\parallel(t_n^\ell)$ using the sequence $(\hat{X}_{\ell,n}^k)_{k=1,\ldots,n}$, constructed with the recursive formula

$$\begin{cases} \hat{X}_{\ell,n}^k = \mathcal{F}(t_n^\ell, t_{n-1}^\ell, \hat{X}_{\ell,n-1}^{k-1}) + \mathcal{G}(t_n^\ell, t_{n-1}^\ell, \hat{X}_{\ell,n-1}^k) - \mathcal{G}(t_n^\ell, t_{n-1}^\ell, \hat{X}_{\ell,n-1}^{k-1}) \\ \hat{X}_{\ell,n}^0 = \mathcal{G}(t_n^\ell, t_{n-1}^\ell, \hat{X}_{\ell,n-1}^0), \quad \hat{X}_{\ell,0}^0 = \hat{x}_\parallel(T_\ell^-). \end{cases} \tag{2.14}$$

In the framework of the Diamond strategy, jumps are computed explicitly and by Proposition 2.2.2, we can determine the number of iterations $k_\ell^{obs} := k$ (required to guarantee a Luenberger's rate of convergence of Algorithm 2.1) a posteriori. Instead, we propose to combine the stopping criterion with estimate (2.11) to derive an a priori upper bound of k_ℓ^{obs}, denoted by k_ℓ, which is given by our next theorem.

Theorem 2.3.2. *We keep the assumptions of Proposition 2.1.2 and Theorem 2.3.1. For a window W_ℓ and $\widetilde{\gamma} > 0$, we define*

$$k_\ell = \begin{cases} \min S_\ell & S_\ell \neq \emptyset \\ k_{\ell-1} & S_\ell = \emptyset \end{cases} \tag{2.15}$$

where

$$\begin{aligned} S_\ell &= \left\{ k \in \mathbb{N}^*, k \leq N-1 : \ \beta^k \sum_{i=0}^{N-k-1} \binom{k+i}{k} \varepsilon^i \leq \frac{\widetilde{\gamma}}{K} \mathrm{e}^{-\mu \ell T} \right\}, \\ K &= \alpha^2 \Big(\frac{1-\varepsilon^N}{1-\varepsilon} \Big). \end{aligned} \tag{2.16}$$

Suppose that we apply Algorithm 2.1 using k_ℓ iterations of the Parareal algorithm, i.e. $h = k_\ell$. Then, the stopping criterion (2.8) is satisfied.

Proof. Combining Theorem 2.3.1 with the definition of k_ℓ in (2.15), we obtain

$$\begin{aligned} \max_{1\leq n\leq N} \left\| J_{\ell,n}^{k_\ell} \right\| &\leq \alpha \beta^{k_\ell} \sum_{i=0}^{n-k_\ell-1} \binom{k_\ell+i}{k_\ell} \varepsilon^i \cdot \max_{1\leq n\leq N} \left\| \hat{X}_{\ell,n}^0 - \hat{x}_\parallel(t_n^\ell) \right\| \\ &\leq \frac{\alpha \widetilde{\gamma}}{K} \mathrm{e}^{-\mu \ell T} \cdot \max_{1\leq n\leq N} \left\| \hat{X}_{\ell,n}^0 - \hat{x}_\parallel(t_n^\ell) \right\| \end{aligned} \tag{2.17}$$

The term $\left\| \hat{X}_{\ell,n}^0 - \hat{x}_\parallel(t_n^\ell) \right\|$ can be bounded as well thanks to Theorem 2.3.1, and then we have

$$\left\| \hat{X}_{\ell,n}^0 - \hat{x}_\parallel(t_n^\ell) \right\| \leq \alpha \sum_{i=0}^{n-1} \varepsilon^i = \alpha \Big(\frac{1-\varepsilon^n}{1-\varepsilon} \Big) \leq \alpha \Big(\frac{1-\varepsilon^N}{1-\varepsilon} \Big),$$

since $\varepsilon < 1$. Thus, from (2.17) and the definition of K in (2.16), we get

$$\max_{1\leq n\leq N} \left\| J_{\ell,n}^{k_\ell} \right\| \leq \widetilde{\gamma} \mathrm{e}^{-\mu \ell T}. \qquad \square$$

Since we introduce an approximation $\hat{X}_{\ell,n}^{k_\ell}$ of the parallel solution $\hat{x}_\parallel(t_n^\ell)$, we provide a bound for the estimation error between the solution $x(t_n^\ell)$ of (2.1) and the parareal sequence.

Corollary 2.3.3. *Given a fixed window W_ℓ, after applying k_ℓ iterations of the Parareal algorithm (2.14) into Algorithm 2.1, we have*

$$\left\| \hat{X}_{\ell,n}^{k_\ell} - x(t_n^\ell) \right\| \leq \left(\widetilde{\gamma}(1 + \mathrm{e}^{-\mu \Delta T}) + \gamma \mathrm{e}^{-\mu \Delta T} \left\| x((0) - \hat{x}(0) \right\| \right) \mathrm{e}^{-\mu \ell T} \tag{2.18}$$

Proof. Given a fixed window W_ℓ, we have

$$\left\|\hat{X}_{\ell,n}^{k_\ell} - x(t_n^\ell)\right\| \leq \max_{1\leq n\leq N} \left\|J_{\ell,n}^{k_\ell}\right\| + \max_{1\leq n\leq N} \left\|\varepsilon_{\|}(t_n^\ell)\right\|$$

Due to Theorem 2.3.2, since Algorithm 2.1 satisfies the stopping criterion (2.8) for k_ℓ iterations of the Parareal algorithm, we use Proposition 2.2.2 to bound both terms. The result follows. □

2.3.3 Complexity analysis

Given a tolerance parameter Tol, we define the efficiency of the algorithm by

$$E = \frac{\tau_s}{N\tau_p} \tag{2.19}$$

where τ_s is the CPU time required to reach the tolerance Tol using a sequential solver. For a parallel solver, we denote this quantity by τ_p, and N represents the number of available processors (and hence, subintervals).

In what follows, we make a distinction between *theoretical* and *observed* efficiency. The former comes from the analysis of our algorithm developped in the previous section, whereas the latter is a result of empirical measurements obtained when using the a posteriori estimator (2.19).

The efficiency of the *Diamond strategy*, combined with the Parareal algorithm, is estimated by our next theorem.

Theorem 2.3.4. *Let* Tol *be the tolerance parameter on the error in Corollary 2.3.3. We define ℓ^* as the number of windows required to reach this tolerance, i.e.*

$$\ell^* = \min\left\{\ell \in \mathbb{N} : \left(\widetilde{\gamma}(1+\mathrm{e}^{-\mu\Delta T}) + \gamma \mathrm{e}^{-\mu\Delta T}\left\|x(0)-\hat{x}(0)\right\|\right)\mathrm{e}^{-\mu\ell T} \leq \mathrm{Tol}\right\}. \tag{2.20}$$

Then, the estimated efficiency of Algorithm 2.1 is given by

$$E^{th} = \frac{\ell^* \tau_{\Delta T}^{\mathcal{F}}}{\tau_{\Delta T}^{\mathcal{F}} + N\tau_{\Delta T}^{\mathcal{G}}} \left(\sum_{\ell=0}^{\ell^*-1} k_\ell\right)^{-1}, \tag{2.21}$$

where $\tau_{\Delta T}^{\mathcal{F}}$, $\tau_{\Delta T}^{\mathcal{G}}$ represents the amount of time spent in solving (2.2) over an interval of size ΔT with $\mathcal{F}$ and $\mathcal{G}$, respectively.

Proof. The decay rate of the Parareal algorithm is provided by Corollary 2.3.3. For a fixed length window T, we define ℓ^* by (2.20). Then, the CPU time for the parallel solver is given by

$$\tau_p = \sum_{\ell=0}^{\ell^*-1}\left[\left(\frac{k_\ell}{N}\right)\tau_{\Delta T}^{\mathcal{F}}\frac{T}{\Delta T} + k_\ell \tau_{\Delta T}^{\mathcal{G}}\frac{T}{\Delta T}\right] = \left(\tau_{\Delta T}^{\mathcal{F}} + N\tau_{\Delta T}^{\mathcal{G}}\right)\sum_{\ell=0}^{\ell^*-1} k_\ell$$

On the other hand, we consider a sequential solver on the interval $[0, \ell^* T]$, which requires $\frac{\ell^* T}{\Delta T}$ iterations. Assuming $\mathcal{F}$ as the solver, its CPU time is given by

$$\tau_s = \tau_{\Delta T}^{\mathcal{F}} \cdot \frac{\ell^* T}{\Delta T} = \tau_{\Delta T}^{\mathcal{F}} \cdot \ell^* N.$$

Finally, from Definition (2.19) of efficiency we get (2.21). □

2.4 Numerical experiments

The present section is devoted to some numerical experiments for the Luenberger observer. For this purpose, we use

$$A = \begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix}, \; B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \; C = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \; u(t) = 3 + 0.5 \sin(0.75t).$$

We remark that the initial condition on System (2.1) is unknown, but we perform the experiments with $x(0) = (0\ 0)^\top$. We then construct the observer $\hat{x}(t)$ by setting as initial condition $\hat{x}(0) = (2\ 1)^\top$ and the eigenvalues of $A - LC$. For the latter, we consider $\{-0.8, -1\}$ and $\{-0.2, -0.25\}$ as possible choices.

To introduce the parareal procedure, given N available processors, we set

$$T = 1, \; \delta T = \Delta T = \frac{T}{N}, \; \text{Tol} = 10^{-8},$$

where δT denotes the time step associated with $\mathcal{G}$, chosen as a one step solver for the sake of simplicity. We use the Backward Euler method to define both propagators $\mathcal{F}$ and $\mathcal{G}$.

2.4.1 Diagonalized system

We recall that the essential part of Theorem 2.3.1 is the contracting factor ε. For the Luenberger observer (2.2), we have

$$\varepsilon = \max\left\{ \left\| [I - \delta t(A - LC)]^{-\Delta T/\delta t} \right\|, \left\| [I - \Delta T(A - LC)]^{-1} \right\| \right\}.$$

where δt is the time step associated with $\mathcal{F}$, assumed to be constant. Even if we choose the eigenvalues of $A - LC$ to guarantee a decaying rate of convergence, ε is not necessarily smaller than one. For this reason, we consider instead a diagonalized observer

$$\begin{cases} \dot{\hat{z}}(t) = D\hat{z}(t) + V^{-1}(Bu(t) + Ly(t)) \\ \hat{z}(0) = V^{-1}\hat{x}_0 \end{cases} \tag{2.22}$$

where $\hat{z} = V^{-1}\hat{x}$ and $D = V^{-1}(A - LC)V$.

Due to the change of variables, $\gamma = 1$. We determine the constants α, β and ε by

Proposition 2.4.1. *Let $\mathcal{F}$ and $\mathcal{G}$ be defined by the Backward Euler scheme, with time steps δt and δT, respectively. We assume that $\Delta T K \leq 1$ and Equation (2.22) satisfies*

$$M := \sup_{(\hat{z},t)} \left\| D\hat{z} + V^{-1}(Bu(t) + Ly(t)) \right\| < \infty$$
$$K := \max \left\{ \|D\|, \sup_{t>0} \left\| V^{-1}(B\dot{u}(t) + L\dot{y}(t)) \right\| \right\} < \infty.$$

Then, the constants associated with both propagators in Theorem 2.3.2 are given by

$$\alpha = \Delta T^2 \left(\frac{K(M+1)}{2(1-\Delta T K)} \right),$$
$$\beta = \left\| [I - \delta t D]^{-\Delta T/\delta t} - [I - \Delta T D]^{-1} \right\|, \tag{2.23}$$
$$\varepsilon = \max \left\{ \left\| [I - \delta t D]^{-\Delta T/\delta t} \right\|, \left\| [I - \Delta T D]^{-1} \right\| \right\}. \tag{2.24}$$

The proof is standard, but we present it for the sake of completeness.

Proof. Let $\{t_n\}_{n=0}^N$ be a regular partition of the interval $[0, T]$, with $\Delta T = T/N$. Given $\hat{z}_{n-1}$ an approximation of $\hat{z}(t_{n-1})$, we recall that the Backward Euler method applied to (2.22) is given by

$$\frac{\hat{z}_n - \hat{z}_{n-1}}{\Delta T} = f(\hat{z}_n, t_n)$$

where $f(s,t) = Ds + V^{-1}(Bv(t) + Ly(t))$.

Since δt is assumed to be constant, we then define $\mathcal{F}$ by

$$\mathcal{F}(t_n, t_{n-1}, \hat{z}_{n-1}) = (I - \delta t D)^{-\Delta T/\delta t} \left[\delta t V^{-1}(Bv(t_{n-1}) + Ly(t_{n-1})) + \hat{z}_{n-1} \right]$$

and then, a direct calculation yields

$$\mathcal{F}(t_n, t_{n-1}, y) - \mathcal{F}(t_n, t_{n-1}, z) = (I - \delta t D)^{-\Delta T/\delta t}(y - z). \tag{2.25}$$

On the other hand, $\mathcal{G}$ is defined as a one-step solver, which allows us to replace δt by ΔT in the previous expressions to obtain

$$\mathcal{G}(t_n, t_{n-1}, y) - \mathcal{G}(t_n, t_{n-1}, z) = (I - \Delta t D)^{-1}(y - z). \tag{2.26}$$

Hence, Definitions (2.23) and (2.24) of β and ε follow from combining (2.25) and (2.26).

To bound the local truncation error, we proceed as follows. Starting at the exact solution $z_{n-1} = \hat{z}(t_{n-1})$, we define $z_n = \mathcal{G}(t_n, t_{n-1}, z_{n-1})$ and then

$$\begin{aligned}\tau(t_n, z_{n-1}) &= \mathcal{F}(t_n, t_{n-1}, z_{n-1}) - \mathcal{G}(t_n, t_{n-1}, z_{n-1}) \\ &= \hat{z}(t_n) - z_n\end{aligned}$$

since $\mathcal{F}$ is an exact solver. We use that $z_n = \hat{z}(t_{n-1}) + \Delta T f(z_n, t_n)$ and then expand $\hat{z}(t_{n-1})$ around t_n to get

$$\tau(t_n, z_{n-1}) = \Delta T \left(\dot{\hat{z}}(\hat{z}(t_n), t_n) - f(z_n, t_n)\right) - \frac{(\Delta T)^2}{2} \ddot{\hat{z}}(\hat{z}(\xi), \xi) \tag{2.27}$$

where $\xi \in (t_{n-1}, t_n)$. Since $\dot{\hat{z}} = f(z, t)$, we can get rid of the derivatives of z. In particular, the definition of $f(s, t)$ shows that

$$\begin{aligned}\dot{\hat{z}}(\hat{z}(t_n), t_n) - f(z_n, t_n) &= f(\hat{z}(t_n), t_n) - f(z_n, t_n) = D\tau(t_n, z_{n-1}), \\ \ddot{\hat{z}}(\hat{z}(\xi), \xi) &= \frac{df}{dt}(\hat{z}(\xi), \xi) = \frac{\partial f}{\partial s}(\hat{z}(\xi), \xi) \cdot f(\hat{z}(\xi), \xi) + \frac{\partial f}{\partial t}(\hat{z}(\xi), \xi) \\ &= Df(\hat{z}(\xi), \xi) + V^{-1}(B\dot{v}(\xi) + L\dot{y}(\xi)).\end{aligned}$$

Replacing these expressions in (2.27) and rearranging terms yields

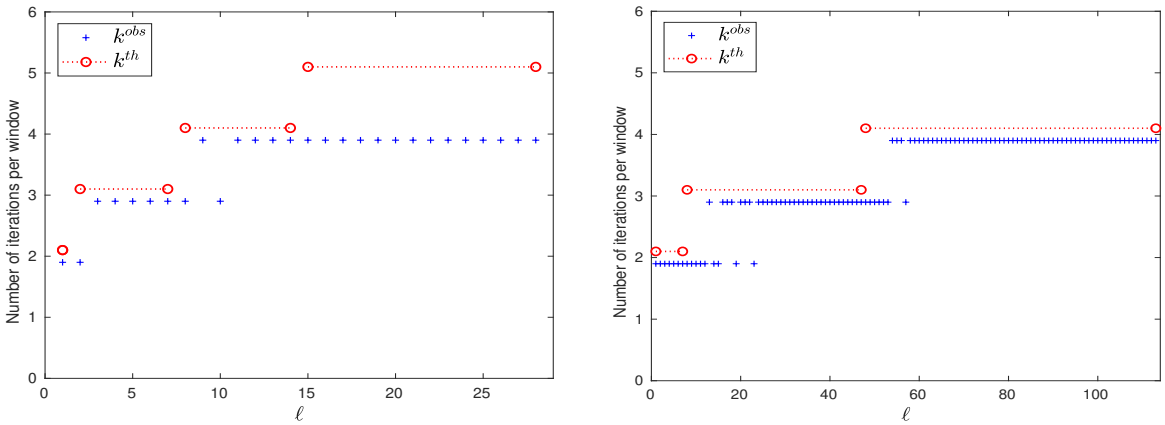
$$\tau(t_n, z_{n-1}) = -\frac{(\Delta T)^2}{2}(I - \Delta T D)^{-1}\Big[Df(\hat{x}(\xi), \xi) + V^{-1}(B\dot{v}(\xi) + L\dot{y}(\xi))\Big].$$

Finally, assuming that $\Delta T K < 1$, we take norms and use the definitions of K and M to obtain α. □

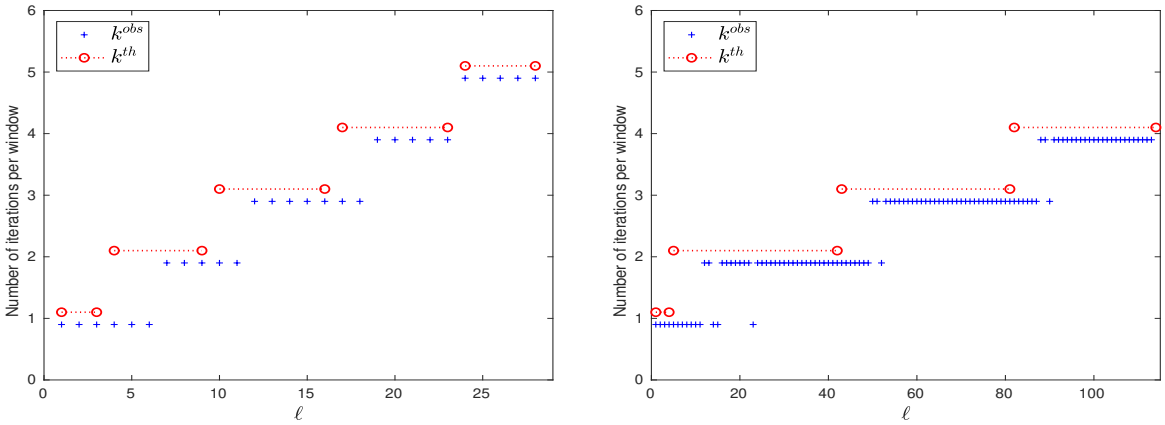
2.4.2 Evolution of k_ℓ

As a first experiment, since the jumps involved in Equation (2.8) allows us to compute the sequence $k^{obs} = \{k_\ell^{obs}\}_\ell$, we propose to compare its behavior with its a priori estimate $k^{th} = \{k_\ell\}_\ell$, provided by Theorem 2.3.2.

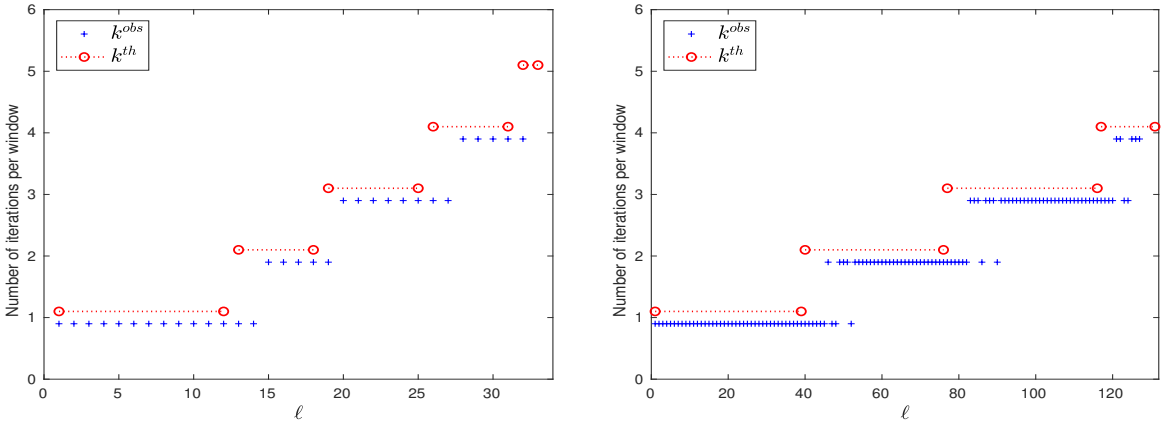
We observe in Figure 2.1 that increasing $\widetilde{\gamma}$ leads to enlarge the number of windows in which the algorithm requires only 1 iteration. This is expected, due to the term $\widetilde{\gamma} e^{-\mu \ell T}$ present in Proposition 2.2.2 and Theorem 2.3.2. We also notice an asymptotical behavior of both sequences in Figure 2.1a, which it is hidden after, because the total number of windows increases at a lower rate.



(a) $\widetilde{\gamma} = 10^{-3}$



(b) $\widetilde{\gamma} = 1$.



(c) $\widetilde{\gamma} = 10^{3}$.

Figure 2.1: Comparison between k^{th} and k^{obs}, for $N = 16$ and $\delta t = \frac{\Delta T}{2^5}$. The eigenvalues of $A - LC$ are $\{-0.8, -1\}$ (left) and $\{-0.2, -0.25\}$ (right).

2.4.3 Observed efficiency

Our second experiment consists in comparing the observed efficiencies for both sequences k^{obs} and k^{th}, using different values of $\widetilde{\gamma}$, N and δt. To evaluate E^{obs}, the execution time for the parallel and sequential solvers was computed with the functions `tic` and `toc` of `MATLAB` (version 9.4.0.813654 (R2018a)).

As we notice previously, increasing $\widetilde{\gamma}$ improves the algorithm performance, but the difference between $E^{obs}(k^{obs})$ and $E^{obs}(k^{th})$ still remains, as observed in Figure 2.2a. In Figure 2.2c, the gap between these values varies slightly, showing that δt small enough does not affect the efficiency significantly. Increasing the number of processors N makes this difference smaller and also improves the efficiency of the algorithm, as shown in Figure 2.2b. Another way to narrow this gap is choosing smaller eigenvalues for $A - LC$. As Figure 2.2 suggests, the comparison between $\{-0.8, -1\}$ and $\{-0.2, -0.25\}$ shows that $E^{obs}(k^{th})$ increases, whereas $E^{obs}(k^{obs})$ becomes smaller.

Figure 2.2 also shows that the observed efficiencies satisfy

$$E^{obs}(k^{th}) \leq E^{obs}(k^{obs}),$$

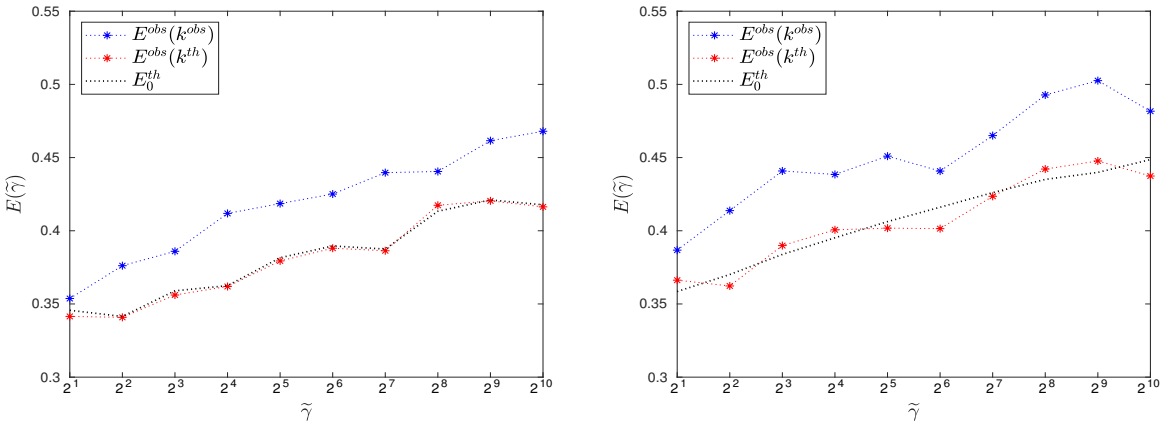
which is simply because the sequence k^{th} underperforms k^{obs}. However, in Figure 2.2b (left) we note at some point the opposite behavior, which could be explained as follows: since β decreases dramatically as N increases, it leads to underestimate the number of iterations. This suggests a relation of the sort

$$\widetilde{\gamma} \leq f(\mu, \Delta T, \delta t).$$

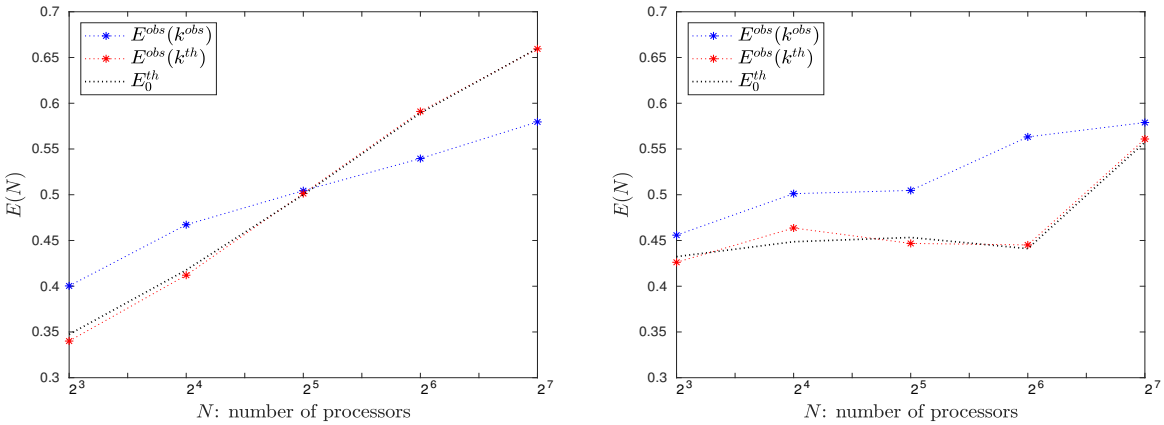
Finally, we recall that k^{th} is useful to determine the theoretical efficiency (2.21). Assuming that $\tau^{\mathcal{G}}_{\Delta T}$ is negligible, we denote it by

$$E_0^{th} = \ell^* \left(\sum_{\ell=0}^{\ell^*-1} k_\ell \right)^{-1}.$$

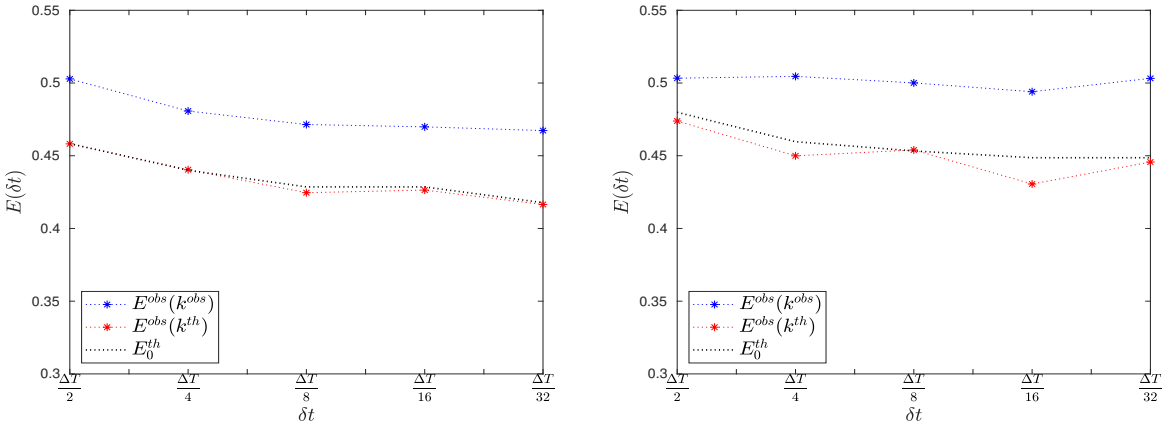
with ℓ^* given by (2.20). We note that this value predicts quite well $E^{obs}(k^{th})$ in all the tests.



(a) $E(\widetilde{\gamma})$, for $N = 16$ and $\delta t = \frac{\Delta T}{2^5}$.



(b) $E(N)$, for $\delta t = \frac{\Delta T}{2^5}$ and $\widetilde{\gamma} = 2^{10}$.



(c) $E(\delta t)$, for $N = 16$ and $\widetilde{\gamma} = 2^{10}$.

Figure 2.2: Comparison between $E^{obs}(k^{obs})$, $E^{obs}(k^{th})$ and E_0^{th}. The eigenvalues of $A - LC$ are $\{-0.8, -1\}$ (left) and $\{-0.2, -0.25\}$ (right).

2.5 Perspectives

The key element of our strategy is the design of a stopping criterion for the parallel-in-time solver that preserves the Luenberger rate of convergence on each assimilation window. Nonetheless, we believe that several questions remains to be answered, starting with using other time-parallelization algorithms, possible extensions to Kalman filters and considering a variable window size.

Another known parallel-in-time algorithm is ParaExp [35]. This method, proposed by Gander and Güttel, decomposes a linear initial-value problem into two subproblems that can be solved in parallel, to then superpose their respective solutions and retrieve the original one. In contrast to Parareal, which often shows a slow convergence when applied to hyperbolic systems, it is well suited for solving these kinds of problem. Its main feature is the approximation of the matrix exponential [70] and consequently, to combine it with our strategy, we propose to determine the number of terms required for a given approximation on each window.

On the other hand, an extension to a stochastic framework should consider a continuous Kalman filter, which is known for being an optimal a posteriori estimator, in the sense that minimizes the variance of the state estimation error. However, an explicit upper bound for this quantity is not often provided. A second point to take into account is that time parallel algorithms are designed for deterministic differential equations and not stochastic ones. The parallelization of the latter was first addressed by Bal [7], followed by applications to e.g. chemical kinetics [29] or finance [76].

Finally, the present work proposes to use an assimilation window of fixed lenght and then determine the number of iterations. The opposite idea, i.e. fixing the latter in a window and determining its size, could be useful when a large number of processors is available.

Chapter 3

Bathymetry optimization

The current chapter focuses on the determination of a bathymetry from an optimization perspective. We consider a PDE-constrained optimization problem in which the wave motion, dependent on the bathymetry and modeled by the weak formulation of the Helmholtz equation, acts as constraint. The cost functional is assumed to be general.

We begin by deriving the Helmholtz equation from the Navier stokes system. Afterwards, a C^0-bound for its weak solution is obtained. Concerning the optimization problem, we study its well-posedness and continuity of the control-to-state mapping. The discrete optimization problem is also addressed, studying the convergence to the discrete optimal solution as well as the convergence of a finite element approximation. We illustrate our results by solving numerically two examples that describe wave damping and bathymetry reconstruction.

This is a joint work with Pierre-Henri Cocquet (Université de Pau et des Pays de l'Adour) and Julien Salomon (ANGE, INRIA Paris).

3.1 Derivation of the wave model

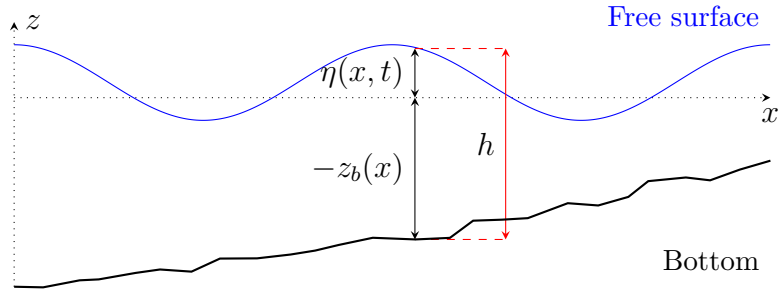
We start from the Navier-Stokes equations to derive the governing PDE. However, due to its complexity, we introduce two approximations [59]: a small relative depth (*Long wave theory*) combined with an infinitesimal wave amplitude (*Small amplitude wave theory*). An asymptotic analysis on the relative depth shows that the vertical component of the depth-averaged velocity is negligible, obtaining the Saint-Venant equations. After neglecting its convective inertia terms and linearizing around the sea level, it results a wave equation which depends on the bathymetry. Since a variable sea bottom can be seen as an obstacle, we reformulate the equation as a *Scattering problem* involving the Helmholtz equation.

3.1.1 From Navier-Stokes system to Saint-Venant equations

For $t \geq 0$, we define the time-dependent region

$$\Omega_t = \{(x, z) \in \Omega \times \mathbb{R} \mid -z_b(x) \leq z \leq \eta(x, t)\}$$

where Ω is a bounded open set with Lipschitz boundary, $\eta(x,t)$ represents the water level and $-z_b(x)$ is the bathymetry (also called bottom topography), a time independent and negative function. The water height is denoted by $h = \eta + z_b$.



In what follows, we consider an incompressible fluid of constant density (assumed to be equal to 1), governed by the Navier-Stokes system

$$\begin{cases} \dfrac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\,\mathbf{u} = \operatorname{div}(\sigma_T) + \mathbf{g} & \text{in } \Omega_t, \\ \operatorname{div}(\mathbf{u}) = 0 & \text{in } \Omega_t, \\ \mathbf{u} = \mathbf{u}_0 & \text{in } \Omega_0, \end{cases} \tag{3.1}$$

where $\mathbf{u} = (u,v,w)^\top$ denotes the velocity of the fluid, $\mathbf{g} = (0,0,-g)^\top$ is the gravity and σ_T is the total stress tensor, given by

$$\sigma_T = -p\mathbb{I} + \mu\left(\nabla\mathbf{u} + \nabla\mathbf{u}^\top\right)$$

with p the pressure and μ the coefficient of viscosity.

To complete (3.1), we require suitable boundary conditions. Given the outward normals

$$n_s = \frac{1}{\sqrt{1+|\nabla\eta|^2}}\begin{pmatrix}-\nabla\eta\\ 1\end{pmatrix},\quad n_b = \frac{1}{\sqrt{1+|\nabla z_b|^2}}\begin{pmatrix}\nabla z_b\\ 1\end{pmatrix},$$

to the free surface and bottom, respectively, we recall that the velocity of the two must be equal to that of the fluid:

$$\begin{cases} \dfrac{\partial\eta}{\partial t} - \mathbf{u}\cdot n_s = 0 & \text{on } (x,\eta(x,t),t), \\ \mathbf{u}\cdot n_b = 0 & \text{on } (x,-z_b(x),t). \end{cases} \tag{3.2}$$

On the other hand, the stress at the free surface is continuous, whereas at the bottom we assume a no-slip condition

$$\begin{cases} \sigma_T\cdot n_s = -p_a n_s & \text{on } (x,\eta(x,t),t), \\ (\sigma_T n_b)\cdot t_b = 0 & \text{on } (x,-z_b(x),t), \end{cases} \tag{3.3}$$

with p_a the atmospheric pressure and t_b an unitary tangent vector to n_b.

A long wave theory approach

For the sake of completeness and following the standard procedure described in [16, 40, 80], we derive the Saint-Venant equations from the Navier-Stokes system. For simplicity of presentation, system (3.1) is restricted to two dimensions, but a more detailed derivation of the three-dimensional case can be found in [27].

Since our analysis focuses on the shallow water regime, we introduce the parameter

$$\varepsilon = \frac{H}{L},$$

where H denotes the relative depth and L is the characteristic dimension along the horizontal axis. The importance of the nonlinear terms is represented by the ratio

$$\delta = \frac{A}{H},$$

with A the maximum vertical amplitude.

We use the change of variables

$$x' = \frac{x}{L},\ z' = \frac{z}{H},\ t' = \frac{C_0}{L}t,$$

and

$$u' = \frac{u}{\delta C_0},\ w' = \frac{w}{\delta\varepsilon C_0},\ \eta' = \frac{\eta}{A},\ z_b' = \frac{z_b}{H},\ p' = \frac{p}{gH},$$

where $C_0 = \sqrt{gH}$ is the characteristic dimension for the horizontal velocity. Assuming the viscosity and atmospheric pressure constants, we define their respective dimensionless versions by

$$\mu' = \frac{\mu}{C_0 L},\ p_a' = \frac{p_a}{gH}.$$

Dropping primes after rescaling, the dimensionless system (3.1) reads

$$\begin{aligned}\delta\frac{\partial u}{\partial t} + \delta^2\left(u\frac{\partial u}{\partial x} + w\frac{\partial u}{\partial z}\right) = &-\frac{\partial p}{\partial x} + 2\delta\frac{\partial}{\partial x}\left(\mu\frac{\partial u}{\partial x}\right), \\ &+ \delta\frac{\partial}{\partial z}\left(\mu\Big(\frac{1}{\varepsilon^2}\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x}\Big)\right)\end{aligned} \tag{3.4}$$

$$\begin{aligned}\varepsilon^2\delta\left(\frac{\partial w}{\partial t} + \delta\Big(u\frac{\partial w}{\partial x} + w\frac{\partial w}{\partial z}\Big)\right) = &-\frac{\partial p}{\partial z} - 1 \\ &+ \delta\frac{\partial}{\partial x}\left(\mu\Big(\frac{\partial u}{\partial z} + \varepsilon^2\frac{\partial w}{\partial x}\Big)\right) + 2\delta\frac{\partial}{\partial z}\left(\mu\frac{\partial w}{\partial z}\right),\end{aligned} \tag{3.5}$$

$$\frac{\partial u}{\partial x} + \frac{\partial w}{\partial z} = 0. \tag{3.6}$$

The boundary conditions in (3.2) remains similar and reads

$$\begin{cases} -\delta u \dfrac{\partial \eta}{\partial x} + w = \dfrac{\partial \eta}{\partial t} \sqrt{1 + (\varepsilon\delta)^2 \left|\dfrac{\partial \eta}{\partial x}\right|^2} & \text{on } (x, \delta\eta(x,t), t), \\ u \dfrac{\partial z_b}{\partial x} + w = 0 & \text{on } (x, -z_b(x), t). \end{cases} \tag{3.7}$$

However, the rescaled boundary conditions in (3.3) are now given by

$$\left(p - 2\delta\mu \frac{\partial u}{\partial x}\right) \frac{\partial \eta}{\partial x} + \mu \left(\frac{1}{\varepsilon^2} \frac{\partial u}{\partial z} + \frac{\partial w}{\partial x}\right) = p_a \frac{\partial \eta}{\partial x} \qquad \text{on } (x, \delta\eta(x,t), t), \tag{3.8}$$

$$\delta^2 \mu \left(\frac{\partial u}{\partial z} + \varepsilon^2 \frac{\partial w}{\partial x}\right) \frac{\partial \eta}{\partial x} + \left(p - 2\delta\mu \frac{\partial w}{\partial z}\right) = p_a \qquad \text{on } (x, \delta\eta(x,t), t), \tag{3.9}$$

and at the bottom $(x, -z_b(x), t)$:

$$\begin{aligned} &\varepsilon \left(p - 2\delta\mu \frac{\partial u}{\partial x}\right) \frac{\partial z_b}{\partial x} + \delta\mu \left(\frac{1}{\varepsilon} \frac{\partial u}{\partial z} + \varepsilon \frac{\partial w}{\partial x}\right) \\ &\quad - \delta\mu \left(\frac{\partial u}{\partial z} + \varepsilon^2 \frac{\partial w}{\partial x}\right) \left(\frac{\partial z_b}{\partial x}\right)^2 + \varepsilon \left(2\delta\mu \frac{\partial w}{\partial z} - p\right) \frac{\partial z_b}{\partial x} = 0. \end{aligned} \tag{3.10}$$

To derive the Saint-Venant equations, we use an asymptotic analysis in ε. In addition, we assume a small viscosity coefficient

$$\mu = \varepsilon\mu_0.$$

A first simplification of the system consists in deriving an explicit expression for p, known as the *hydrostatic pressure*. After rearranging the terms of order ε^2 in (3.5) and integrating in the vertical direction, we get

$$\begin{aligned} p(x,z,t) = \mathcal{O}(\varepsilon^2\delta) + (\delta\eta - z) + \varepsilon\delta\mu_0 \left(\frac{\partial u}{\partial x} + 2\frac{\partial w}{\partial z} - \frac{\partial u}{\partial x}(x, \delta\eta, t)\right) \\ + p(x, \delta\eta, t) - 2\varepsilon\delta\mu_0 \frac{\partial w}{\partial z}(x, \delta\eta, t). \end{aligned} \tag{3.11}$$

To compute explicitly the last term, we combine (3.8) with (3.9) to obtain

$$\begin{aligned} p(x, \delta\eta, t) - 2\varepsilon\delta\mu_0 \frac{\partial w}{\partial z}(x, \delta\eta, t) = p_a \left(1 - (\varepsilon\delta)^2 \left(\frac{\partial \eta}{\partial x}\right)^2\right) \\ + (\varepsilon\delta)^2 \left(p - 2\varepsilon\mu_0 \frac{\partial u}{\partial x}(x, \delta\eta, t)\right) \left(\frac{\partial \eta}{\partial x}\right)^2, \end{aligned}$$

and then, replacing this quantity into (3.11) yields

$$p(x,z,t) = (\delta\eta - z) + p_a + \mathcal{O}(\varepsilon\delta). \tag{3.12}$$

As a second approximation, we integrate vertically Equations (3.6) and (3.4). We denote the depth-averaged velocity by

$$\overline{u}(x,t) = \frac{1}{h_\delta(x,t)} \int_{-z_b}^{\delta\eta} u(x,z,t)dz,$$

where $h_\delta = \delta\eta + z_b$. Due to the Leibnitz integral rule and the boundary conditions in (3.7), integrating the mass equation (3.6) gives

$$\begin{aligned}
\int_{-z_b}^{\delta\eta} \left(\frac{\partial u}{\partial x} + \frac{\partial w}{\partial z}\right) dz = 0 \\
\frac{\partial}{\partial x}\left(\int_{-z_b}^{\delta\eta} u dz\right) - \delta u(x,\delta\eta,t)\frac{\partial \eta}{\partial x} - u(x,-z_b,t)\frac{\partial z_b}{\partial x} + w(x,\delta\eta,t) - w(x,-z_b,t) = 0 \\
\frac{\partial \eta}{\partial t}\sqrt{1+(\varepsilon\delta)^2\left|\frac{\partial \eta}{\partial x}\right|^2} + \frac{\partial (h_\delta \overline{u})}{\partial x} = 0.
\end{aligned}$$

To treat the momentum equation (3.4), we notice that Equation (3.6) allows us to rewrite the convective acceleration terms as

$$u\frac{\partial u}{\partial x} + w\frac{\partial u}{\partial z} = \frac{\partial u^2}{\partial x} + \frac{\partial uw}{\partial z}.$$

Its integration, combined with the boundary conditions in (3.7), leads to

$$\begin{aligned}
\int_{-z_b}^{\delta\eta}\left(u\frac{\partial u}{\partial x} + w\frac{\partial u}{\partial z}\right)dz &= \frac{\partial}{\partial x}\left(\int_{-z_b}^{\delta\eta} u^2 dz\right) - \delta u^2(x,\delta\eta,t)\frac{\partial \eta}{\partial x} - u^2(x,-z_b,t)\frac{\partial z_b}{\partial x} \\
&\quad + u(x,\delta\eta,t)\cdot w(x,\delta\eta,t) - u(x,-z_b,t)\cdot w(x,-z_b,t) \\
&= \frac{\partial (h_\delta \overline{u^2})}{\partial x} + u(x,\delta\eta,t)\frac{\partial \eta}{\partial t}\sqrt{1+(\varepsilon\delta)^2\left|\frac{\partial \eta}{\partial x}\right|^2},
\end{aligned}$$

and then, the vertical integration of the left-hand side of (3.4) brings

$$\begin{aligned}
\int_{-z_b}^{\delta\eta}\left[\delta\frac{\partial u}{\partial t} + \delta^2\left(u\frac{\partial u}{\partial x} + w\frac{\partial u}{\partial z}\right)\right]dz &= \delta\frac{\partial (h_\delta\overline{u})}{\partial t} + \delta^2\frac{\partial (h_\delta \overline{u^2})}{\partial x} \\
&\quad + \delta^2 u(x,\delta\eta,t)\frac{\partial \eta}{\partial t}\left(\sqrt{1+(\varepsilon\delta)^2\left|\frac{\partial \eta}{\partial x}\right|^2} - 1\right).
\end{aligned}$$

To deal with the term $h_\delta\overline{u^2}$, we start from (3.12) which shows that $\frac{\partial p}{\partial x} = \mathcal{O}(\delta)$. Plugging this expression into (3.4) yields

$$\frac{\partial^2 u}{\partial z^2} = \mathcal{O}(\varepsilon).$$

From boundary conditions (3.8) and (3.10) we obtain

$$\frac{\partial u}{\partial z}(x,\delta\eta,t)=\mathcal{O}(\varepsilon^2),\ \frac{\partial u}{\partial z}(x,z_b,t)=\mathcal{O}(\varepsilon).$$

Consequently, $u(x,z,t)=u(x,0,t)+\mathcal{O}(\varepsilon)$ and then $u(x,z,t)-\overline{u}(x,t)=\mathcal{O}(\varepsilon)$. Hence, we have the approximation

$$h_\delta\overline{u^2}=h_\delta\overline{u}^2+\int_{-z_b}^{\delta\eta}(\overline{u}-u)^2dz=h_\delta\overline{u}^2+\mathcal{O}(\varepsilon^2)$$

and finally

$$\int_{-z_b}^{\delta\eta}\left[\delta\frac{\partial u}{\partial t}+\delta^2\left(u\frac{\partial u}{\partial x}+w\frac{\partial u}{\partial z}\right)\right]dz=\delta\frac{\partial(h_\delta\overline{u})}{\partial t}+\delta^2\frac{\partial(h_\delta\overline{u}^2)}{\partial x}+\mathcal{O}(\varepsilon^2\delta^2)\qquad(3.13)$$
$$+\delta^2u(x,\delta\eta,t)\frac{\partial\eta}{\partial t}\left(\sqrt{1+(\varepsilon\delta)^2\left|\frac{\partial\eta}{\partial x}\right|^2}-1\right).$$

We then integrate the right-hand side of Equation (3.4)

$$\int_{-z_b}^{\delta\eta}\left[-\frac{\partial p}{\partial x}+\delta\frac{\mu_0}{\varepsilon}\frac{\partial}{\partial z}\left(\frac{\partial u}{\partial z}\right)+\varepsilon\delta\mu_0\left(2\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right)+\frac{\partial}{\partial z}\left(\frac{\partial w}{\partial x}\right)\right)\right]dz$$
$$=-\delta h_\delta\frac{\partial\eta}{\partial x}+\mathcal{O}(\varepsilon\delta)+\delta\left[\frac{\mu_0}{\varepsilon}\frac{\partial u}{\partial z}(x,\delta\eta,t)-\frac{\mu_0}{\varepsilon}\frac{\partial u}{\partial z}(x,-z_b,t)\right].$$

Combining this expression with (3.13), we get the vertical integration of the momentum equation.

In summary, we have the system

$$\frac{\partial\eta}{\partial t}\sqrt{1+(\varepsilon\delta)^2\left|\frac{\partial\eta}{\partial x}\right|^2}+\frac{\partial(h_\delta\overline{u})}{\partial x}=0\qquad(3.14)$$

$$\frac{\partial(h_\delta\overline{u})}{\partial t}+\delta\frac{\partial(h_\delta\overline{u}^2)}{\partial x}=-h_\delta\frac{\partial\eta}{\partial x}+\mathcal{O}(\varepsilon)\qquad(3.15)$$
$$+\left[\frac{\mu_0}{\varepsilon}\frac{\partial u}{\partial z}(x,\delta\eta,t)-\frac{\mu_0}{\varepsilon}\frac{\partial u}{\partial z}(x,-z_b,t)\right]$$
$$+\delta u(x,\delta\eta,t)\frac{\partial\eta}{\partial t}\left(\sqrt{1+(\varepsilon\delta)^2\left|\frac{\partial\eta}{\partial x}\right|^2}-1\right).$$

If $\delta=\mathcal{O}(1)$ and $\varepsilon\to0$, then we recover the classical derivation of the one-dimensional Saint-Venant equations. The convergence of (3.15) is guaranteed by the boundary equations (3.8) and (3.10), from which we get

$$\frac{\mu_0}{\varepsilon}\frac{\partial u}{\partial z}(x,\delta\eta,t)=\mathcal{O}(\varepsilon\delta),\ \frac{\mu_0}{\varepsilon}\frac{\partial u}{\partial z}(x,-z_b,t)=\mathcal{O}(\varepsilon).$$

Small amplitudes

With respect to the classical Saint-Venant formulation, passing to the limit $\delta \to 0$ is equivalent to neglect the convective acceleration terms and linearize the system (3.14-3.15) around the sea level $\eta = 0$. In order to do so, we rewrite the derivatives as

$$\frac{\partial(h_\delta \overline{u})}{\partial t} = h_\delta \frac{\partial \overline{u}}{\partial t} + \delta \frac{\partial \eta}{\partial t}\overline{u}, \quad \frac{\partial(h_\delta \overline{u})}{\partial x} = \delta \frac{\partial(\eta \overline{u})}{\partial x} + \frac{\partial(z_b \overline{u})}{\partial x},$$

and then, taking $\varepsilon, \delta \to 0$ in (3.14-3.15) yields

$$\begin{cases} \dfrac{\partial \eta}{\partial t} + \dfrac{\partial(z_b \overline{u})}{\partial x} = 0, \\ -\dfrac{\partial(z_b \overline{u})}{\partial t} + z_b \dfrac{\partial \eta}{\partial x} = 0. \end{cases}$$

Finally, after deriving the first equation with respect to t and replacing the second into the new expression, we obtain the wave equation for a variable bathymetry.

All the previous computations hold for the three-dimensional system (3.1). In this case, we obtain

$$\frac{\partial^2 \eta}{\partial t^2} - \operatorname{div}(g z_b \nabla \eta) = 0. \tag{3.16}$$

3.1.2 Helmholtz formulation

Equation (3.16) defines a time-harmonic field, whose solution has the form $\eta(x,t) = \mathrm{Re}\{\psi_{tot}(x)\mathrm{e}^{-\mathrm{i}\omega t}\}$, where the amplitude ψ_{tot} satisfies

$$\omega^2 \psi_{tot} + \operatorname{div}(g z_b \nabla \psi_{tot}) = 0. \tag{3.17}$$

We wish to rewrite the equation above as a scattering problem. Since a variable bottom $z_b(x) := z_0 + \delta z_b(x)$ (with z_0 a constant describing a flat bathymetry and δz_b a perturbation term) can be considered as an obstacle, we thus assume that δz_b has a compact support in Ω and ψ_{tot} satisfies the so-called Sommerfeld radiation condition. In a bounded domain as Ω, we impose the latter thanks to an impedance boundary condition (also known as first-order absorbing boundary condition), which ensures the existence and uniqueness of the solution [75, p.108]. We then reformulate (3.17) as

$$\begin{cases} \operatorname{div}((1+q)\nabla \psi_{tot}) + k_0^2 \psi_{tot} = 0 & \text{in } \Omega, \\ \nabla(\psi_{tot} - \psi_0) \cdot \hat{n} - \mathrm{i}k_0(\psi_{tot} - \psi_0) = 0 & \text{on } \partial\Omega, \end{cases} \tag{3.18}$$

where $q(x) := \frac{\delta z_b(x)}{z_0}$ is compactly supported in Ω, $k_0 := \frac{\omega}{\sqrt{g z_0}}$, $\hat{n}$ is the unit normal to $\partial\Omega$ and $\psi_0(x) = \mathrm{e}^{\mathrm{i}k_0 x \cdot \vec{d}}$ is an incident plane wave propagating in the direction $\vec{d}$ (such that $|\vec{d}| = 1$).

Decomposing the total wave as $\psi_{tot} = \psi_0 + \psi_{sc}$, where ψ_{sc} represents an unknown scattered wave, we obtain the Helmholtz formulation

$$\begin{cases} \operatorname{div}\left((1+q)\nabla\psi_{sc}\right) + k_0^2\psi_{sc} = -\operatorname{div}\left(q\nabla\psi_0\right) & \text{in } \Omega, \\ \nabla\psi_{sc}\cdot\hat{n} - \mathrm{i}k_0\psi_{sc} = 0 & \text{on } \partial\Omega. \end{cases} \tag{3.19}$$

Its structure will be useful to prove the existence of a minimizer for a PDE-constrained functional, as discussed in the next section.

3.2 Description of the optimization problem

We are interested in studying the problem of a cost functional constrained by the weak formulation of a Helmholtz equation. The latter intends to generalize the equations considered so far, whereas the former indirectly affects the choice of the set of admissible controls. These can be discontinuous since they are included in the space of functions of bounded variations. In this framework, we treat the continuity and regularity of the associated control-to-state mapping, and the existence of an optimal solution to the optimization problem.

3.2.1 Weak formulation

Let $\Omega \subset \mathbb{R}^2$ be a bounded open set with Lipschitz boundary. We consider the following general Helmholtz equation

$$\begin{cases} -\operatorname{div}\left((1+q)\nabla\psi\right) - k_0^2\psi = \operatorname{div}\left(q\nabla\psi_0\right) & \text{in } \Omega, \\ (1+q)\nabla\psi\cdot\hat{n} - \mathrm{i}k_0\psi = g - q\nabla\psi_0\cdot\hat{n} & \text{on } \partial\Omega, \end{cases} \tag{3.20}$$

where g is a source term. We assume that $q \in L^\infty(\Omega)$ and that there exists $\alpha > 0$ such that

$$\text{a.e. } x \in \Omega,\ 1 + q(x) \geq \alpha. \tag{3.21}$$

Remark 3.2.1. *Here we have generalized the models described in the previous section: if q has a fixed compact support in Ω, we have that the total wave ψ_{tot} satisfying (3.18) is a solution to (3.20) with $g = \nabla\psi_0\cdot\hat{n} - \mathrm{i}k_0\psi_0$ and no volumic right-hand side; whereas the scattered wave ψ_{sc} satisfying (3.19) is a solution to (3.20) with $g = 0$. All the results obtained in this broader setting still hold true for both problems.*

A weak formulation for (3.20) is given by

$$a(q;\psi,\phi) = b(q;\phi),\ \forall\phi \in H^1(\Omega), \tag{3.22}$$

where

$$\begin{aligned} a(q;\psi,\phi) &:= \int_\Omega \left((1+q)\nabla\psi\cdot\nabla\overline{\phi} - k_0^2\psi\overline{\phi}\right)dx - \mathrm{i}k_0\int_{\partial\Omega}\psi\overline{\phi}\,d\sigma, \\ b(q;\phi) &:= -\int_\Omega q\nabla\psi_0\cdot\nabla\overline{\phi}\,dx + \langle g,\overline{\phi}\rangle_{H^{-1/2},H^{1/2}}. \end{aligned}$$

Note that, thanks to Cauchy-Schwarz inequality, the sesquilinear form a is continuous

$$|a(q;\psi,\phi)| \leq C(\Omega,q,\alpha)(1+\|q\|_{L^\infty(\Omega)})\,\|\psi\|_{1,k_0}\,\|\phi\|_{1,k_0}\,,$$
$$\|\psi\|^2_{1,k_0} := k_0^2\,\|\psi\|^2_{L^2(\Omega)} + \alpha\,\|\nabla\psi\|^2_{L^2(\Omega)}\,,$$

where $C(\Omega,q,\alpha)>0$ is a generic constant. In addition, taking $\phi=\psi$ in the definition of a, it satisfies a Garding inequality

$$\mathrm{Re}\{a(q;\psi,\psi)\} + 2k_0^2\,\|\psi\|^2_{L^2(\Omega)} \geq \|\psi\|^2_{1,k_0}\,, \tag{3.23}$$

and the well-posedness of Problem (3.22) follows from the Fredholm Alternative. Uniqueness holds for any $q\in L^\infty(\Omega)$ satisfying (3.21) owning to [46, Theorems 2.1, 2.4].

3.2.2 Continuous optimization problem

We are interested in solving the next PDE-constrained optimization problem

$$\begin{aligned} &\min_{(q,\psi)\in U_\Lambda\times H^1(\Omega)} J(q,\psi) \\ &\text{s.t.} \qquad (3.22). \end{aligned} \tag{3.24}$$

We now define the set of admissible q. We wish to find optimal q that can have discontinuities and we thus cannot look for q in some Sobolev spaces that are continuously embedded into $C^0(\overline{\Omega})$, even if such regularity is useful for proving existence of a minimizer (see e.g. [8, Chapter VI], [13, Theorem 4.1]). To be able to find optimal q satisfying (3.21) and having possible discontinuities, we follow [21] and introduce the set

$$U_\Lambda = \{q\in BV(\Omega)\ \mid \alpha-1\leq q(x)\leq\Lambda\ a.e.\ x\in\Omega\}\,.$$

Above $\Lambda\geq\max\{\alpha-1,0\}$ and $BV(\Omega)$ is the set of functions with bounded variations [5] whose distributional gradient belong to the set $\mathcal{M}_\mathrm{b}(\Omega,\mathbb{R}^N)$ of bounded Radon measures. Note that piecewise constant functions over Ω belong to U_Λ.

Some useful properties of $BV(\Omega)$ can be found in [5] and are recalled below for the sake of completeness. This is a Banach space for the norm (see [5, p.120, Proposition 3.2])

$$\|q\|_{BV(\Omega)} := \|q\|_{L^1(\Omega)} + |Dq|(\Omega),$$

where D is the distributional gradient and

$$|Dq|(\Omega) = \sup\left\{\int_\Omega q\,\mathrm{div}\,(\varphi)\;dx\ \middle|\ \varphi\in\mathcal{C}^1_\mathrm{c}(\Omega,\mathbb{R}^2)\ \text{and}\ \|\varphi\|_{L^\infty(\Omega)}\leq 1\right\},$$

is the variation of q (see [5, p.119, Definition 3.4]).

The weak* convergence in $BV(\Omega)$, denoted by

$$q_n \rightharpoonup q, \text{ weak}^* \text{ in } BV(\Omega),$$

means that

$$q_n \to q \text{ in } L^1(\Omega) \text{ and } Dq_n \rightharpoonup Dq \text{ in } \mathcal{M}_{\mathrm{b}}(\Omega, \mathbb{R}^N).$$

Also, in a two-dimensional setting, the continuous embedding $BV(\Omega) \subset L^1(\Omega)$ is compact. We finally recall that the application $q \in BV(\Omega) \mapsto |Dq|(\Omega) \in \mathbb{R}^+$ is lower semi-continuous with respect to the weak* topology of BV. Hence, for any sequence $q_n \rightharpoonup q$ in $BV(\Omega)$, we have

$$|Dq|(\Omega) \leq \liminf_{n \to +\infty} |Dq_n|(\Omega).$$

The set U_Λ is a closed, weakly* closed and convex subset of $BV(\Omega)$. However, since its elements are not necessarily bounded in the BV-norm, in order to prove the existence of a minimizer to Problem (3.24) we need to add a penalizing distributional gradient term to the cost functional $J(q, \psi)$. To avoid this situation, we introduce the set of admissible parameters

$$U_{\Lambda,\kappa} = \{q \in U_\Lambda \mid |Dq|(\Omega) \leq \kappa\}$$

which also possesses the aforementioned properties. The choice between these two sets also affects the convergence analysis of the discrete optimization problem, topic discussed in Section 3.4.

Remark 3.2.2. *In this chapter, we are interested in computing either the total wave satisfying (3.18) or the scattered wave solution to Equation (3.19). Since this require to work with q having a fixed compact support in Ω, we also introduce the following set of admissible parameters*

$$\widetilde{U}_\varepsilon := \{q \in U \ |q(x) = 0 \text{ for a.e } x \in \mathcal{O}_\varepsilon\}, \ \mathcal{O}_\varepsilon = \{x \in \Omega \mid \operatorname{dist}(x, \partial\Omega) \leq \varepsilon\},$$

which is a set of bounded functions with bounded variations that have a fixed support in Ω. We emphasize that it is a convex, closed and weak-$*$ closed subset of $BV(\Omega)$. As a result, all the theorems we are going to prove also hold for this set of admissible parameters.

3.2.3 Continuity of the control-to-state mapping

This section is devoted to prove the continuity of the application $q \in U \mapsto \psi(q) \in H^1(\Omega)$ where $\psi(q)$ satisfies Problem (3.22). We assume that $U \subset BV(\Omega)$ is weakly* closed and

$$\forall q \in U, \text{ a.e. } x \in \Omega, \ \alpha - 1 \leq q(x) \leq \Lambda.$$

Note that both U_Λ, $U_{\Lambda,\kappa}$ and $\widetilde{U}_\varepsilon$ (see Remark 3.2.2) also satisfy these two assumptions. The next result consider the dependance of the stability constant with respect to the optimization parameter q.

Theorem 3.2.3. *Assume that $q \in U$. Then there exists a constant $C_s(k_0) > 0$ that does not depend on q such that*

$$\|\psi\|_{1,k_0} \leq C_s(k_0) \sup_{\|\phi\|_{1,k_0}=1} |a(q;\psi,\phi)|, \tag{3.25}$$

where the constant $C_s(k_0) > 0$ only depend on the wavenumber and on Ω. In addition, the solution to (3.22) satisfies the bound

$$\|\psi\|_{1,k_0} \leq C_s(k_0) C(\Omega) \max\{k_0^{-1}, \alpha^{-1/2}\} \left(\|q\|_{L^\infty(\Omega)} \|\nabla \psi_0\|_{L^2(\Omega)} + \|g\|_{H^{-1/2}(\partial\Omega)} \right), \tag{3.26}$$

where $C(\Omega) > 0$ only depends on the domain.

Proof. We recall that the existence and uniqueness of solutions to Problem (3.22) was already stated in Subsection 3.2.1.

The proof of (3.25) proceed by contradiction assuming the above inequality is false. Therefore, we suppose there exist sequences $(q_n)_n \subset U$ and $(\psi_n)_n \subset H^1(\Omega)$ such that $\|q_n\|_{BV(\Omega)} \leq M$, $\|\psi_n\|_{1,k_0} = 1$ and

$$\lim_{n\to+\infty} \sup_{\|\phi\|_{1,k_0}=1} |a(q_n;\psi_n,\phi)| = 0. \tag{3.27}$$

The compactness of the embeddings $BV(\Omega) \subset L^1(\Omega)$ and $H^1(\Omega) \subset L^2(\Omega)$ yields the existence of a subsequence (still denoted (q_n, ψ_n)) such that

$$\psi_n \rightharpoonup \psi_\infty \text{ in } H^1(\Omega),\ \psi_n \to \psi_\infty \text{ in } L^2(\Omega) \text{ and } q_n \to q_\infty \in U \text{in } L^1(\Omega). \tag{3.28}$$

Compactness of the trace operator gives that $\lim_{n\to+\infty} \psi_n|_{\partial\Omega} = \psi_\infty|_{\partial\Omega}$ strongly in $L^2(\partial\Omega)$ and thus, from (3.28), we get

$$\lim_{n\to+\infty} \int_\Omega k_0^2 \psi_n \overline{\phi}\, dx + \mathrm{i}k_0 \int_{\partial\Omega} \psi_n \overline{\phi}\, d\sigma = \int_\Omega k_0^2 \psi_\infty \overline{\phi}\, dx + \mathrm{i}k_0 \int_{\partial\Omega} \psi_\infty \overline{\phi}\, d\sigma,\ \forall\ v \in H^1(\Omega),$$
$$\lim_{n\to+\infty} \int_\Omega \nabla\psi_n \cdot \nabla\overline{\phi}\, dx = \int_\Omega \nabla\psi_\infty \cdot \nabla\overline{\phi}\, dx.$$

It only remains to pass to the limit in the term involving q_n. We start from

$$(q_n \nabla\psi_n, \nabla\overline{\phi})_{L^2(\Omega)} - (q_\infty \nabla\psi_\infty, \nabla\overline{\phi})_{L^2(\Omega)} = ((q_n - q_\infty)\nabla\psi_n, \nabla\overline{\phi})_{L^2(\Omega)} + (q_\infty \nabla(\psi_n - \psi_\infty), \nabla\overline{\phi})_{L^2(\Omega)},$$

and use Cauchy-Schwarz inequality to get

$$\begin{aligned}
&\int_\Omega q_n \nabla\psi_n \cdot \nabla\overline{\phi}\, dx - \int_\Omega q_\infty \nabla\psi_\infty \cdot \nabla\overline{\phi}\, dx \\
&\qquad \leq \left|((q_n - q_\infty)\nabla\psi_n, \nabla\overline{\phi})_{L^2(\Omega)}\right| + \left|(q_\infty \nabla(\psi_n - \psi_\infty), \nabla\overline{\phi})_{L^2(\Omega)}\right| \\
&\qquad \leq \left\|\sqrt{|q_n - q_\infty|}\nabla\phi\right\|_{L^2(\Omega)} \left\|\sqrt{|q_n - q_\infty|}\nabla\psi_n\right\|_{L^2(\Omega)} \\
&\qquad\quad + \left|(q_\infty \nabla(\psi_n - \psi_\infty), \nabla\overline{\phi})_{L^2(\Omega)}\right| \\
&\qquad \leq 2\frac{\sqrt{\Lambda}}{\sqrt{\alpha}} \|\psi_n\|_{1,k_0} \left\|\sqrt{|q_n - q_\infty|}\nabla\phi\right\|_{L^2(\Omega)} + \left|(\nabla(\psi_n - \psi_\infty), q_\infty \nabla\overline{\phi})_{L^2(\Omega)}\right|.
\end{aligned}$$

The right term above goes to 0 owning to $q_\infty \in L^\infty(\Omega)$ and (3.28). For the other term, since $q_n \to q_\infty$ strongly in L^1, we can extract another subsequence $(q_{n_k})_k$ such that $q_{n_k} \to q_\infty$ pointwise almost everywhere in Ω. Also, $\sqrt{|q_n - q_\infty|}|\nabla\phi|^2 \leq 2\sqrt{\Lambda}|\nabla\phi|^2 \in L^1(\Omega)$ and the Lebesgue's Dominated Convergence Theorem then yields

$$\lim_{k\to+\infty} \left\| \sqrt{|q_{n_k} - q_\infty|}\nabla\phi \right\|_{L^2(\Omega)} = 0.$$

This gives that (see also [21, Equation (2.4)])

$$\lim_{k\to+\infty} (q_{n_k}\nabla\psi_{n_k}, \nabla\overline{\phi})_{L^2(\Omega)} = (q_\infty\nabla\psi_\infty, \nabla\overline{\phi})_{L^2(\Omega)}, \ \forall\phi \in H^1(\Omega). \tag{3.29}$$

Finally, gathering (3.29) together with (3.27) yields

$$0 = \lim_{k\to+\infty} a(q_{n_k}; \psi_{n_k}, \phi) = a(q_\infty, \psi_\infty, \phi), \ \forall\phi \in H^1(\Omega),$$

and the uniqueness result [46, Theorems 2.1, 2.4] shows that $\psi_\infty = 0$ thus the whole sequence actually converges to 0. To get our contradiction, it remains to show that $\|\nabla\psi_n\|_{L^2(\Omega)}$ converges to 0 as well. From the Garding inequality (3.23), we have

$$\|\psi_n\|^2_{1,k_0} \leq \mathrm{Re}\{a(q_n; \psi_n, \psi_n)\} + 2k_0^2 \|\psi_n\|^2_{L^2(\Omega)} \xrightarrow[n\to+\infty]{} 0,$$

where we used (3.27) and the strong L^2 convergence of ψ_n towards $\psi_\infty = 0$. Finally, we get $\lim_{n\to+\infty} \|\psi_n\|_{1,k_0} = 0$ which contradicts $\|\psi_n\|_{1,k_0} = 1$ and gives the desired result.

Applying then (3.25) to the solution to (3.22) finally yields

$$\begin{aligned}
\|\psi\|_{1,k_0} &\leq C_\mathrm{s}(k_0) \sup_{\|\phi\|_{1,k_0}=1} |a(q;\psi,\phi)| \leq C_\mathrm{s}(k_0) \sup_{\|\phi\|_{1,k_0}=1} |b(q;\phi)| \\
&\leq C_\mathrm{s}(k_0) \sup_{\|\phi\|_{1,k_0}=1} \left(\|q\|_{L^\infty(\Omega)} \|\nabla\psi_0\|_{L^2(\Omega)} \|\nabla\phi\|_{L^2(\Omega)} + \|g\|_{H^{-1/2}(\partial\Omega)} \|\phi\|_{H^{1/2}(\partial\Omega)} \right) \\
&\leq C_\mathrm{s}(k_0) C(\Omega) \max\{k_0^{-1}, \alpha^{-1/2}\} \left(\|q\|_{L^\infty(\Omega)} \|\nabla\psi_0\|_{L^2(\Omega)} + \|g\|_{H^{-1/2}(\partial\Omega)} \right),
\end{aligned}$$

where $C(\Omega) > 0$ comes from the trace inequality. □

Remark 3.2.4. *Let us consider a more general version of Problem (3.20), given by*

$$\begin{cases} -\mathrm{div}\left((1+q)\nabla\psi\right) - k_0^2\psi = F & \text{in } \Omega, \\ (1+q)\nabla\psi\cdot\hat{n} - ik_0\psi = G & \text{on } \partial\Omega. \end{cases}$$

We emphasize that the estimation of the stability constant $C_\mathrm{s}(k_0)$ with respect to the wavenumber have been obtained for $(F,G) \in L^2(\Omega)\times L^2(\partial\Omega)$ for $q = 0$ in [49] and for $q \in \mathrm{Lip}(\Omega)$ satisfying (3.21) in[11, 46, 45]. Since their proofs rely on Green, Rellich and Morawetz identities, they do not extend to the case $(F,G) \in (H^1(\Omega))' \times H^{-1/2}(\partial\Omega)$ but such cases can be tackled as it is done in [31, p.10, Theorem 2.5]. The case of Lipschitz q has been studied in [17]. As a result, the dependance of the stability constant with respect to q, in the case $q \in U$ and $(F,G) \in (H^1(\Omega))' \times H^{-1/2}(\partial\Omega)$, does not seem to have been tackled so far to the best of our knowledge.

Remark 3.2.5 (H^1-bounds for the total and scattered waves)**.** *From Remark 3.2.1, we obtain that the total wave ψ_{tot} and the scattered wave ψ_{sc} are solutions to (3.22), with respective right-hand sides*

$$b_{tot}(q;\phi) = \int_{\partial\Omega} (\nabla\psi_0 \cdot \hat{n} - \mathrm{i}k_0\psi_0)\overline{\phi}\, d\sigma, \quad b_{sc}(q;\phi) = -\int_{\Omega} q\nabla\psi_0 \cdot \nabla\overline{\phi}\, dx.$$

As a result of Theorem 3.2.3 and the continuity of the trace, we have

$$\|\psi_{tot}\|_{1,k_0} \leq C(\Omega)C_{\mathrm{s}}(k_0)k_0 \max\{k_0^{-1}, \alpha^{-1/2}\},$$
$$\|\psi_{sc}\|_{1,k_0} \leq C_{\mathrm{s}}(k_0)\alpha^{-1/2}\, \|q\|_{L^\infty(\Omega)}\, \|\nabla\psi_0\|_{L^2(\Omega)} \leq k_0 C_{\mathrm{s}}(k_0)\alpha^{-1/2}\, \|q\|_{L^\infty(\Omega)} \sqrt{|\Omega|}.$$

We can now prove some regularity for the control-to-state mapping.

Theorem 3.2.6. *Let $(q_n)_n \subset U$ be a sequence satisfying $\|q_n\|_{BV(\Omega)} \leq M$ and whose weak* limit in $BV(\Omega)$ is denoted by q_∞. Let $(\psi(q_n))_n$ be the sequence of weak solution to Problem (3.22). Then $\psi(q_n)$ converges strongly in $H^1(\Omega)$ towards $\psi(q_\infty)$. In other words, the mapping*

$$q \in (U_\Lambda, \text{weak}^*) \mapsto \psi(q) \in (H^1(\Omega), \text{strong}),$$

is continuous.

Proof. Since $\|q_n\|_{BV(\Omega)} \leq M$ and $(q_n)_n \subset U$, there exists q_∞ such that $q_n \rightharpoonup q_\infty$, weak* in $BV(\Omega)$. Using that U is weak* closed, we obtain that $q_\infty \in U$. Note that the sequence $(\psi(q_n))_n$ of solution to Problem (3.22) satisfies estimate (3.26) uniformly with respect to n. As a result, the convergences (3.28) hold and, from (3.29) and the unicity of the limiting problem, we get that $a(q_n;\psi_n,\phi) \to a(q_\infty;\psi_\infty,\phi)$. Since $b(q_n,\phi) \to b(q_\infty,\phi)$ for all $\phi \in H^1(\Omega)$, this proves that $a(q_\infty;\psi_\infty,\phi) = b(q;\phi)$ for all $\phi \in H^1(\Omega)$, $\psi(q_\infty)$ is a weak solution to (3.22) and $\psi(q_n) \rightharpoonup \psi(q_\infty)$ in $H^1(\Omega)$.

We now show that $\psi(q_n) \to \psi(q_\infty)$ strongly in H^1. To see this, we start by noting that, up to extracting a subsequence (still denoted q_n), we can use (3.29) to get that

$$\lim_{n\to+\infty} b(q_n;\psi(q_n)) = b(q_\infty;\psi(q_\infty)).$$

Since $\psi(q_n), \psi(q_\infty)$ satisfy the variational problem (3.22), we infer

$$\lim_{n\to+\infty} a(q_n;\psi(q_n),\psi(q_n)) = a(q_\infty;\psi_\infty,\psi(q_\infty)), \tag{3.30}$$

where the whole sequence actually converges owing to the uniqueness of the limit. Using then that $\psi(q_n) \rightharpoonup \psi(q_\infty)$ in $H^1(\Omega)$ together with (3.30), we get

$$\begin{aligned}\left\|\sqrt{1+q_n}\nabla\psi(q_n)\right\|^2_{L^2(\Omega)} &= a(q_n;\psi(q_n),\psi(q_n)) + k_0\, \|\psi(q_n)\|^2_{L^2(\Omega)} + \mathrm{i}k_0\, \|\psi(q_n)\|^2_{L^2(\partial\Omega)} \\ &\xrightarrow[n\to+\infty]{} a(q_\infty;\psi(q_\infty),\psi(q_\infty)) + k_0\, \|\psi(q_\infty)\|^2_{L^2(\Omega)} + \mathrm{i}k_0\, \|\psi(q_\infty)\|^2_{L^2(\partial\Omega)} \\ &= \left\|\sqrt{1+q_\infty}\nabla\psi(q_\infty)\right\|^2_{L^2(\Omega)}.\end{aligned}$$

To show that $\lim_{n\to+\infty} \|\nabla\psi(q_n)\|^2_{L^2(\Omega)} = \|\nabla\psi(q_\infty)\|^2_{L^2(\Omega)}$, note that

$$\nabla\psi(q_n) = \frac{\sqrt{1+q_n}\nabla\psi(q_n)}{\sqrt{1+q_n}}.$$

Using the same arguments as those to prove (3.29), we have a subsequence (same notation used) such that $q_n \to q_\infty$ pointwise a.e. in Ω and thus $\sqrt{1+q_n}^{-1} \to \sqrt{1+q_\infty}^{-1}$ pointwise a.e. in Ω. Due to Lebesgue's Dominated Convergence Theorem and $\sqrt{1+q_n}\nabla\psi(q_n) \to \sqrt{1+q_\infty}\nabla\psi(q_\infty)$ strongly in $L^2(\Omega)$, we have

$$\nabla\psi(q_n) = \frac{\sqrt{1+q_n}\nabla\psi(q_n)}{\sqrt{1+q_n}} \to \frac{\sqrt{1+q_\infty}\nabla\psi(q_\infty)}{\sqrt{1+q_\infty}} = \nabla\psi(q_\infty) \text{ strong in } L^2(\Omega).$$

The latter, together with Theorem 3.2.6, shows that $\psi(q_n) \to \psi(q_\infty)$ strongly in H^1. □

3.2.4 Existence of optimal solution

We are now in position to prove the existence of a minimizer to Problem (3.24) for $U = U_\Lambda$.

Theorem 3.2.7. *Assume that the cost function $(q,\psi) \in U_\Lambda \mapsto J(q,\psi) \in \mathbb{R}$ satisfies:*

(A1) There exists $\beta > 0$ such that

$$J(q,\psi) = J_0(q,\psi) + \beta|Dq|(\Omega).$$

(A2) $\forall (q,\psi) \in U_\Lambda \times H^1(\Omega)$, $J_0(q,\psi) \geq m > -\infty$.

(A3) $(q,\psi) \mapsto J_0(q,\psi)$ is lower-semi-continuous with respect to the (weak,weak) topology of $BV(\Omega) \times H^1(\Omega)$.*

Then the optimization problem (3.24) has at least one optimal solution in $U_\Lambda \times H^1(\Omega)$.

Proof. The existence of a minimizer to Problem (3.24) can be obtained with standard techniques by combining Theorem 3.2.6 with weak-compactness arguments as done in [21, Lemma 2.1], [13, Theorem 4.1] or [48, Theorem 1]. We still give the proof for the sake of completeness.

We introduce the following set

$$\mathcal{A} = \left\{ (q,\psi) \in U_\Lambda \times H^1(\Omega) \;\middle|\; a(q;\psi,\phi) = b(q;\phi)\ \forall \phi \in H^1(\Omega) \right\}.$$

The existence and uniqueness of solution to Problem (3.22) ensure that $\mathcal{A}$ is non-empty. In addition, combining assumptions $(A1)$ and $(A2)$ we have that $J(q,\psi)$ is

bounded from below on $\mathcal{A}$. We thus have a minimizing sequence $(q_n, \psi_n) \in \mathcal{A}$ such that

$$\lim_{n\to+\infty} J(q_n, \psi_n) = \inf_{(q,\psi)\in\mathcal{A}} J(q, \psi).$$

Theorem 3.2.3 and $(A1)$ then gives that the sequence $(q_n, \psi_n) \in BV(\Omega) \times H^1(\Omega)$ is uniformly bounded with respect to n and thus admits a subsequence that converges towards (q^*, ψ^*) in the (weak*,weak) topology of $BV(\Omega)\times H^1(\Omega)$. Using now Theorem 3.2.6 and the weak* lower semi-continuity of $q \mapsto |Dq|(\Omega)$, we end up with $(q^*, \psi^*) \in \mathcal{A}$ and

$$J(q^*, \psi^*) \leq \liminf_{n\to+\infty} J(q_n, \psi_n) = \inf_{(q,\psi)\in\mathcal{A}} J(q, \psi). \qquad \square$$

It is worth noting that β has been introduced only to obtain a uniform bound in the BV-norm for the minimizing sequence.

When $U = U_{\Lambda,\kappa}$, we note that any $q \in U_{\Lambda,\kappa}$ is actually bounded in BV since

$$\|q\|_{BV(\Omega)} \leq 2\max(\Lambda, \kappa, |\alpha - 1|)$$

With this property at hand, we can get a similar result to Theorem 3.2.7 without adding a penalization term in the cost function, hence $\beta = 0$.

Theorem 3.2.8. *Assume that the cost function $(q, \psi) \in U_{\Lambda,\kappa} \mapsto J(q, u) \in \mathbb{R}$ satisfies $(A2) - (A3)$ given in Theorem 3.2.7 and that $\beta = 0$. Then the optimization problem (3.24) with $U = U_{\Lambda,\kappa}$ has at least one optimal solution.*

Proof. We introduce the following non-empty set

$$\mathcal{A} = \left\{(q, \psi) \in U_{\Lambda,\kappa} \times H^1(\Omega) \mid a(q; \psi, \phi) = b(q; \phi)\ \forall \phi \in H^1(\Omega)\right\}.$$

From $(A2)$, $J(q, \psi)$ is bounded from below on $\mathcal{A}$. We thus have a minimizing sequence $(q_n, \psi_n) \in \mathcal{A}$ such that

$$\lim_{n\to+\infty} J(q_n, \psi_n) = \inf_{(q,\psi)\in\mathcal{A}} J(q, \psi).$$

Since $(q_n)_n \subset U_{\Lambda,\kappa}$, it satisfies $\|q_n\|_{BV(\Omega)} \leq 2\max(\Lambda, \kappa, |\alpha - 1|)$ and thus admits a convergent subsequence toward some $q \in U_{\Lambda,\kappa}$. Theorem 3.2.6 then gives that $\psi(q_n) \to \psi(q)$ strongly in $H^1(\Omega)$ and the proof can be finished as the proof of Theorem 3.2.7. $\square$

3.3 Boundedness/Continuity of solution to Helmholtz problem

We prove in this section that, even if the parameter q is not smooth enough for the solution to (3.20) to be in $H^s(\Omega)$ for some $s > 1$, we can still have continuous solution. In order to prove such regularity for ψ, we are going to rely on the De Giorgi-Nash-Moser theory [41, Section 8.5], [57, Sections 3.13, 7.2] and more precisely on [73, Proposition 3.6] which reads

Theorem 3.3.1. *Consider the elliptic problem with inhomogeneous Neumann boundary condition*

$$\begin{cases} \mathcal{L}v := \operatorname{div}(A(x)\nabla v) = f_0 - \sum_{j=1}^{N} \dfrac{\partial f_j}{\partial x_j}, \\ \nabla v \cdot \hat{n} = h + \sum_{j=1}^{N} f_j n_j, \end{cases} \tag{3.31}$$

where $A \in L^\infty(\Omega, \mathbb{R}^{N\times N})$ satisfy the standard ellipticity condition $A(x)\xi\cdot\xi \geq \gamma|\xi|^2$ for a.e. $x \in \Omega$. Let $p > N$ and assume that $f_0 \in L^{p/2}(\Omega)$, $f_j \in L^p(\Omega)$ for all $j = 1, \cdots, N$ and $h \in L^{p-1}(\partial\Omega)$. Then the weak solution v to (3.31) satisfies

$$\|v\|_{C^0(\Omega)} \leq C(N, p, \Omega, \gamma) \left(\|v\|_{L^2(\Omega)} + \|f_0\|_{L^{p/2}(\Omega)} + \sum_{j=1}^{N} \|f_j\|_{L^p(\Omega)} + \|h\|_{L^{p-1}(\partial\Omega)} \right).$$

3.3.1 C^0-bound for the general Helmholtz problem

Using Theorem 3.3.1, we can prove some L^∞ bound for the weak solution to Helmholtz equation with bounded coefficients.

Theorem 3.3.2. *Assume that $q \in L^\infty(\Omega)$ and satisfies (3.21) and $g \in L^2(\partial\Omega)$. Then the solution to Problem (3.22) satisfies*

$$\|\psi\|_{C^0(\Omega)} \leq \widetilde{C}(\Omega)\widetilde{C}_{\mathrm{s}}(k_0, \alpha) \left(\|q\|_{L^\infty(\Omega)} \|\nabla\psi_0\|_{L^\infty(\Omega)} + \|g\|_{L^2(\partial\Omega)} \right), \tag{3.32}$$

where

$$\widetilde{C}_{\mathrm{s}}(k_0, \alpha) = 1 + \left((1 + k_0^2)k_0^{-1} + \alpha^{-1/2} \right) \max\{k_0^{-1}, \alpha^{-1/2}\} C_{\mathrm{s}}(k_0),$$

and $\widetilde{C}(\Omega) > 0$ does not depend on k nor q.

Proof. We cannot readily apply Theorem 3.3.1 to the weak solution of Problem (3.20) since the latter involves a complex valued operator. We therefore consider the Problem satisfied by $\nu = \operatorname{Re}\{u\}$ and $\zeta = \operatorname{Im}\{u\}$ which is given by

$$\begin{cases} -\operatorname{div}((1+q)\nabla\nu) - k_0^2\nu = \operatorname{div}(q\nabla \operatorname{Re}\{\psi_0\}) & \text{in } \Omega, \\ -\operatorname{div}((1+q)\nabla\zeta) - k_0^2\zeta = \operatorname{div}(q\nabla \operatorname{Im}\{\psi_0\}) & \text{in } \Omega, \\ (1+q)\nabla\nu\cdot\hat{n} = \operatorname{Re}\{g\} - k_0\zeta - q\nabla \operatorname{Re}\{\psi_0\}\cdot\hat{n}, & \text{on } \partial\Omega, \\ (1+q)\nabla\zeta\cdot\hat{n} = \operatorname{Im}\{g\} + k_0\nu - q\nabla \operatorname{Im}\{\psi_0\}\cdot\hat{n} & \text{on } \partial\Omega. \end{cases} \tag{3.33}$$

Since Problem (3.33) is equivalent to Problem (3.20), we get that the weak solution $(\nu, \zeta) \in H^1(\Omega)$ to (3.33) satisfies the inequality (3.26). Assuming that $g \in L^2(\partial\Omega)$ and using the continuous Sobolev embedding $H^1(\Omega) \subset L^6(\Omega)$, the (compact) embedding $H^{1/2}(\partial\Omega) \subset L^2(\partial\Omega)$, that $q \in L^\infty(\Omega)$ satisfies (3.21) and the fact that ψ_0 is smooth we get the next regularities

$$f_{0,1} = k_0^2\nu \in L^6(\Omega),\ f_{j,1} = q\frac{\partial \operatorname{Re}\{\psi_0\}}{\partial x_j} \in L^\infty(\Omega),\ h_1 = \operatorname{Re}\{g\} - k_0\zeta \in L^2(\partial\Omega),$$
$$f_{0,2} = k_0^2\zeta \in L^6(\Omega),\ f_{j,2} = q\frac{\partial \operatorname{Im}\{\psi_0\}}{\partial x_j} \in L^\infty(\Omega),\ h_2 = \operatorname{Im}\{g\} + k_0\nu \in L^2(\partial\Omega).$$

Applying now Theorem 3.3.1 to (3.33) twice with $p = 3$ and $N = 2$, we get C^0 bounds for ν and ζ

$$\|\nu\|_{C^0(\Omega)} \leq C(2,3,\Omega,\gamma)\left(\|\nu\|_{L^2(\Omega)} + \|f_{0,1}\|_{L^{3/2}(\Omega)} + \sum_{j=1}^{2}\|f_{j,1}\|_{L^3(\Omega)} + \|h_1\|_{L^2(\partial\Omega)}\right),$$
$$\|\zeta\|_{C^0(\Omega)} \leq C(2,3,\Omega,\gamma)\left(\|\zeta\|_{L^2(\Omega)} + \|f_{0,2}\|_{L^{3/2}(\Omega)} + \sum_{j=1}^{2}\|f_{j,2}\|_{L^3(\Omega)} + \|h_2\|_{L^2(\partial\Omega)}\right).$$

Some computations with the Holder and multiplicative trace inequalities then give

$$\begin{aligned}
(\|\nu\|_{L^2(\Omega)} + \|\zeta\|_{L^2(\Omega)}) &\leq 2\|\psi\|_{L^2(\Omega)},\\
\|f_{0,1}\|_{L^{3/2}(\Omega)} + \|f_{0,2}\|_{L^{3/2}(\Omega)} &\leq k_0^2\|\psi\|_{L^{3/2}(\Omega)} \leq |\Omega|^{1/6}k_0^2\|\psi\|_{L^2(\Omega)},\\
\|f_{j,l}\|_{L^3(\Omega)} &\leq \|q\|_{L^\infty(\Omega)}\|\nabla\psi_0\|_{L^\infty(\Omega)},\ j = 1,2,\\
\|h_1\|_{L^2(\partial\Omega)} + \|h_2\|_{L^2(\partial\Omega)} &\leq \|g\|_{L^2(\partial\Omega)} + k_0\|\psi\|_{L^2(\partial\Omega)},\\
&\leq \|g\|_{L^2(\partial\Omega)} + k_0C(\Omega)\sqrt{\|\psi\|_{L^2(\Omega)}\|\psi\|_{H^1(\Omega)}}.
\end{aligned}$$

Using then Young inequality yields

$$\begin{aligned}
k_0\sqrt{\|\psi\|_{L^2(\Omega)}\|\psi\|_{H^1(\Omega)}} &\leq C\left(\|\psi\|_{H^1(\Omega)} + k_0^2\|\psi\|_{L^2(\Omega)}\right)\\
&\leq C\left(\|\nabla\psi\|_{L^2(\Omega)} + (1+k_0^2)\|\psi\|_{L^2(\Omega)}\right)
\end{aligned}$$

where $C > 0$ is a generic constant. We obtain the bound

$$\begin{aligned}
\|\psi\|_{C^0(\Omega)} &= \|\nu\|_{C^0(\Omega)} + \|\zeta\|_{C^0(\Omega)}\\
&\leq \widetilde{C}(\Omega)\Big(\left(1+k_0^2\right)\|\psi\|_{L^2(\Omega)} + \|\nabla\psi\|_{L^2(\Omega)}\\
&\qquad + \|q\|_{L^\infty(\Omega)}\|\nabla\psi_0\|_{L^\infty(\Omega)} + \|g\|_{L^2(\partial\Omega)}\Big).
\end{aligned}$$

Using the definition of $\|\psi\|_{1,k_0}$ on the estimate above, we have

$$\begin{aligned}
\|\psi\|_{C^0(\Omega)} \leq \widetilde{C}(\Omega)\Big(&\left((1+k_0^2)k_0^{-1} + \alpha^{-1/2}\right)\|\psi\|_{1,k_0}\\
&+ \|q\|_{L^\infty(\Omega)}\|\nabla\psi_0\|_{L^\infty(\Omega)} + \|g\|_{L^2(\partial\Omega)}\Big)
\end{aligned} \tag{3.34}$$

To apply the a priori estimate (3.26), we recall that the $H^{-1/2}$ norm can be replaced by a L^2 norm (since $g \in L^2(\partial\Omega)$) and then,

$$\begin{aligned}\|\psi\|_{1,k_0} &\leq C(\Omega)\max\{k_0^{-1},\alpha^{-1/2}\}C_{\mathrm{s}}(k_0)\left(\|q\|_{L^\infty(\Omega)}\|\nabla\psi_0\|_{L^2(\Omega)}+\|g\|_{L^2(\partial\Omega)}\right)\\ &\leq C(\Omega)\max\{k_0^{-1},\alpha^{-1/2}\}C_{\mathrm{s}}(k_0)\max\{1,\sqrt{|\Omega|}\}\\ &\qquad\times\left(\|q\|_{L^\infty(\Omega)}\|\nabla\psi_0\|_{L^\infty(\Omega)}+\|g\|_{L^2(\partial\Omega)}\right)\end{aligned}$$

Finally, combining the last expression with Equation (3.34), we obtain that the weak solution to the Helmholtz equation satisfies

$$\begin{aligned}\|\psi\|_{C^0(\Omega)} \leq \widetilde{C}(\Omega)\left(1+\left((1+k_0^2)k_0^{-1}+\alpha^{-1/2}\right)\max\{k_0^{-1},\alpha^{-1/2}\}C_{\mathrm{s}}(k_0)\right)\\ \times\left(\|q\|_{L^\infty(\Omega)}\|\nabla\psi_0\|_{L^\infty(\Omega)}+\|g\|_{L^2(\partial\Omega)}\right),\end{aligned}$$

where $\widetilde{C}(\Omega) > 0$. □

Remark 3.3.3. *1. For the one-dimensional Helmholtz problem, the a priori estimate (3.26) and the continuous embedding $H^1(I) \subset C^0(I)$ directly gives the continuity of u over I*

$$\|\psi\|_{C^0(I)} \leq C\|\psi\|_{1,k_0} \leq C(k)\left(\|q\|_{L^\infty(\Omega)}\|\nabla\psi_0\|_{L^\infty(\Omega)}+\|g\|_{H^{-1/2}(\partial\Omega)}\right).$$

It is worth noting that we do not need to assume that $g \in L^2(\partial\Omega)$.

2. For the two-dimensional Helmholtz problem with $q = 0$, we can get the above $\mathcal{C}^0$ estimate from the embedding $H^2(\Omega) \hookrightarrow \mathcal{C}^0(\overline{\Omega})$ since

$$\|\psi\|_{C^0(\Omega)} \leq C\|\psi\|_{H^2(\Omega)},$$

for a generic constant C. We can then see that the estimate (3.32) actually has the same dependance with respect to k_0 as the H^2-estimate [49, p.677, Proposition 3.6].

3.3.2 C^0-bounds for the total and scattered waves

Thanks to Remark 3.2.1 and following the proof of Theorem 3.3.2, these bounds can be roughly obtained by setting $g = \nabla\psi_0 \cdot \hat{n} - \mathrm{i}k_0\psi_0$ and omitting the L^∞-norms in (3.34) for the total wave ψ_{tot}, and simply by setting $g = 0$ in the case the scattered wave ψ_{sc}. Using after the H^1-bounds from Remark 3.2.5, we actually get

$$\begin{aligned}\|\psi_{tot}\|_{\mathcal{C}^0(\Omega)} &\leq \widetilde{C}(\Omega)k_0\left(\left((1+k_0^2)k_0^{-1}+\alpha^{-1/2}\right)\max\{k_0^{-1},\alpha^{-1/2}\}C_{\mathrm{s}}(k_0)+1\right)\\ \|\psi_{sc}\|_{\mathcal{C}^0(\Omega)} &\leq \widetilde{C}(\Omega)k_0\left(\left((1+k_0^2)k_0^{-1}+\alpha^{-1/2}\right)\alpha^{-1/2}C_{\mathrm{s}}(k_0)+1\right)\|q\|_{L^\infty(\Omega)}.\end{aligned}$$

We emphasize that the previous estimates show that the scattered wave ψ_{sc} vanishes in Ω if $q \to 0$. This is expected since, if $q = 0$, there is no obstacle to scatter the incident wave, which amount to saying that $\psi_{tot} = \psi_0$.

3.4 Discrete optimization problem

This section is devoted to the finite element discretization of the optimization problem (3.24). We consider a quasi-uniform family of triangulations (see [30, p.76, Definition 1.140]) $\{\mathcal{T}_h\}_{h>0}$ of Ω and the corresponding finite element spaces

$$\mathcal{V}_h = \left\{\phi_h \in \mathcal{C}(\overline{\Omega}) \mid \phi_h|_T \in \mathbb{P}_1(T),\ \forall T \in \mathcal{T}_h\right\}.$$

Note that thanks to Theorem 3.3.2, the solution to the general Helmholtz equation (3.20) is continuous, which motivates to use continuous piecewise linear finite elements. We are going to look for discrete optimal design that belong to some finite element spaces $\mathcal{K}_h$ and we thus introduce the following set of discrete admissible parameters

$$U_h = U \cap \mathcal{K}_h.$$

The full discretization of the optimization problem (3.24) then reads

$$\text{Find } q_h^* \in U_h \text{ such that } \widetilde{J}(q_h^*) \leq \widetilde{J}(q_h),\ \forall q_h \in U_h, \tag{3.35}$$

where $\widetilde{J}(q_h) = J(q_h, \psi_h(q_h))$ is the reduced cost-functional and $\psi_h := \psi_h(q_h) \in \mathcal{V}_h$ satisfies the discrete Helmholtz problem

$$a(q_h; \psi_h, \phi_h) = b(q_h; \phi_h),\ \forall \phi_h \in \mathcal{V}_h. \tag{3.36}$$

The existence of solution to Problem (3.36) follows from uniqueness since we are in a finite dimensional setting [46, Theorems 2.1, 2.4].

Before giving the definition of $\mathcal{K}_h$, we would like to discuss briefly the strategy for proving that the discrete optimal solution converges toward the continuous ones. To achieve this, we need to pass to the limit in inequality (3.35). Since J is only lower-semi-continuous with respect to the weak* topology of BV, we can only pass to the limit on one side of the inequality and the continuity of J is then going to be needed to pass to the limit on the other side to keep this inequality valid as $h \to 0$. We discuss first the case $U = U_\Lambda$ for which Theorem 3.2.7 gives the existence of optimal q but only if $\beta > 0$. Since we have to pass to the limit in (3.35), we need that $\lim_{h\to 0} |Dq_h|(\Omega) = |Dq|(\Omega)$. Since the total variation is only continuous with respect to the strong topology of BV, we have to approximate any $q \in U_\Lambda$ by some $q_h \in U_h$ such that

$$\lim_{h\to 0} \|q - q_h\|_{BV(\Omega)} = 0.$$

However, from [9, p.8, Example 4.1] there is an example of a BV-function v that cannot be approximated by piecewise constant function v_h over a given mesh in such a way that $\lim_{h\to 0} |Dv_h|(\Omega) = |Dv|(\Omega)$. Nevertheless, if one consider an adapted mesh that depends on a given function $v \in BV(\Omega) \cap L^\infty(\Omega)$, we get the existence of piecewise constant function on this specific mesh that strongly converges in BV toward v (see

[18, p.11, Theorem 4.2]). As a result, when considering $U = U_\Lambda$, we use the following discrete set of admissible parameters

$$\mathcal{K}_{h,1} = \{q_h \in L^\infty(\Omega) \mid q_h|_T \in \mathbb{P}_1(T),\ \forall T \in \mathcal{T}_h\}\,.$$

Note that, from Theorem [18, p.10, Theorem 4.1 and Remark 4.2], the set $U_h = U_\Lambda \cap \mathcal{K}_{h,1}$ defined above has the required density property hence motivated its introduction as a discrete set of admissible parameter.

In the case, $U = U_{\Lambda,\kappa}$, we will not need the density of U_h in the strong topology of BV but only for the weak* topology. The discrete set of admissible parameter is then going to be $U_h = U_{\Lambda,\kappa} \cap \mathcal{K}_{h,0}$ with

$$\mathcal{K}_{h,0} = \{q_h \in L^\infty(\Omega) \mid q_h|_T \in \mathbb{P}_0(T),\ \forall T \in \mathcal{T}_h\}\,.$$

We show below the convergence of discrete optimal solution to the continuous one for both cases highlighted above.

3.4.1 Convergence of the Finite element approximation

We prove here some useful approximations results for any U_h defined above. We have the following convergence result whose proof can be found in [31, p.22, Lemma 4.1] (see also [46, p.10, Theorem 4.1]).

Theorem 3.4.1. *Let $q_h \in U_h$ and $\psi(q_h) \in H^1(\Omega)$ be the solution to the variational problem*

$$a(q_h; \psi(q_h), \phi) = b(q_h, \phi),\ \forall \phi \in H^1(\Omega).$$

Let $S^* : (q_h, f) \in U_h \times L^2(\Omega) \mapsto S^*(q_h, f) = \psi^* \in H^1(\Omega)$ be the solution operator associated with the following problem

$$\text{Find } \psi^* \in H^1(\Omega) \text{ such that } a(q_h; \phi, \psi^*) = (\phi, \overline{f})_{L^2(\Omega)},\ \forall \phi \in H^1(\Omega).$$

Denote by C_a the continuity constant of the bilinear form $a(q_h; \cdot, \cdot)$, which does not depend on h since $q_h \in U_h$, and define the adjoint approximation property by

$$\delta(\mathcal{V}_h) := \sup_{f \in L^2(\Omega)} \inf_{\phi_h \in \mathcal{V}_h} \frac{\|S^*(q_h, f) - \phi_h\|_{1,k_0}}{\|f\|_{L^2(\Omega)}}.$$

Assume that the spaces $\mathcal{V}_h$ satisfies

$$2C_a k_0 \delta(\mathcal{V}_h) \leq 1, \tag{3.37}$$

then the solution $\psi_h(q_h)$ to Problem (3.36) satisfies

$$\|\psi(q_h) - \psi_h(q_h)\|_{1,k_0} \leq 2C_a \inf_{\phi_h \in \mathcal{V}_h} \|\psi(q_h) - \phi_h\|_{1,k_0}\,.$$

In the case $q \in \mathcal{C}^{0,1}(\Omega)$ where Ω is a convex Lipschitz domain, the assumption (3.37) has been discussed in [46, p.11, Theorem 4.3] and, broadly speaking, (3.37) holds if $k_0^2 h$ is small enough. Since the proof rely on H^2-regularity for a Poisson problem, we cannot readily extend the argument here since we can only expect to have $\psi \in H^1(\Omega)$ and that S^* also depend on the meshsize. We can still show that (3.37) is satisfied for small enough h.

Lemma 3.4.2. *Assume that $q_h \in U_h$ weak* converges toward $q \in BV(\Omega)$. Then (3.37) is satisfied for small enough h.*

Proof. Note first that Theorem 3.2.6 also holds for the adjoint problem and thus

$$\lim_{h\to 0} \|S^*(q_h, f) - S^*(q, f)\|_{1,k_0} = 0.$$

Using the density of smooth function in H^1 and the properties of the piecewise linear interpolant [30, p.66, Corollary 1.122], we have that

$$\lim_{h\to 0} \left(\sup_{f\in L^2(\Omega)} \inf_{\phi_h \in \mathcal{V}_h} \frac{\|S^*(q, f) - \phi_h\|_{1,k_0}}{\|f\|_{L^2(\Omega)}} \right) = 0,$$

and thus a triangular inequality shows that (3.37) holds for small enough h. □

We can now prove a discrete counterpart to Theorem 3.2.6.

Theorem 3.4.3. *Let $(q_h)_h \subset U_h$ be a sequence satisfying $\|q_h\|_{BV(\Omega)} \leq M$ and whose weak* limit in $BV(\Omega)$ is denoted by q. Let $(\psi_h(q_h))_h$ be the sequence of discrete solutions to Problem (3.36). Then $\psi(q_h)$ converges, as h goes to 0, strongly in $H^1(\Omega)$ towards $\psi(q)$ satisfying Problem (3.22).*

Proof. For h small enough, Lemma 3.4.2 ensures that (3.37) holds and a triangular inequality then yields

$$\begin{aligned}
\|\psi_h(q_h) - \psi(q)\|_{1,k_0} &\leq \|\psi_h(q_h) - \psi(q_h)\|_{1,k_0} + \|\psi(q_h) - \psi(q)\|_{1,k_0} \\
&\leq 2C_a \inf_{\phi_h \in \mathcal{V}_h} \|\psi(q_h) - \phi_h\|_{1,k_0} + \|\psi(q_h) - \psi(q)\|_{1,k_0} \\
&\leq (1 + 2C_a) \|\psi(q_h) - \psi(q)\|_{1,k_0} + 2C_a \inf_{\phi_h \in \mathcal{V}_h} \|\psi(q) - \phi_h\|_{1,k_0}.
\end{aligned}$$

Theorem 3.2.6 gives that the first term above goes to zero as $h \to 0$. For the second one, we can use the density of smooth function in H^1 to get that it goes to zero as well. □

3.4.2 Convergence of the discrete optimal solution

We are now in position to prove the convergence of discrete optimal design toward continuous one in the case

$$U = U_\Lambda, \; U_h = U_\Lambda \cap \mathcal{K}_{h,1}.$$

Hence the set of discrete control is composed of piecewise linear function on $\mathcal{T}_h$.

Theorem 3.4.4. *Assume that* $(A1)-(A2)-(A3)$ *from Theorem 3.2.7 hold and that the cost function* $J_0 : (q,\psi) \in U_\Lambda \times H^1(\Omega) \mapsto J_0(q,\psi) \in \mathbb{R}$ *is continuous with respect to the* (weak*, strong) *topology of* $BV(\Omega) \times H^1(\Omega)$. *Let* $(q_h^*, \psi_h(q_h^*)) \in U_{\Lambda,h} \times \mathcal{V}_h$ *be an optimal pair of (3.35). Then the sequence* $(q_h^*)_h \subset U_\Lambda$ *is bounded and there exists* $q^* \in U_\Lambda$ *such that* $q_h^* \rightharpoonup q^*$ *weakly** *in* $BV(\Omega)$, $\psi(q_h^*) \to \psi(q^*)$ *strongly in* $H^1(\Omega)$ *and*

$$\widetilde{J}(q^*) \leq \widetilde{J}(q), \ \forall q \in U_\Lambda.$$

Hence any accumulation point of $(q_h^*, \psi_h(q_h^*))$ *is an optimal pair for Problem* (3.24).

Proof. Let $q_\Lambda \in U_{\Lambda,h}$ be given as

$$q_\Lambda(x) = \Lambda, \ \forall x \in \Omega.$$

Then $Dq_\Lambda = 0$. Since $\psi_h(q_\Lambda)$ is well-defined and converges toward $\psi(q_\Lambda)$ strongly in H^1 (see Theorem 3.4.4), we have that

$$\widetilde{J}(q_\Lambda) = J(q_\Lambda, \psi_h(q_\Lambda)) = J_0(q_\Lambda, \psi_h(q_\Lambda)) \xrightarrow[h\to 0]{} J_0(q_\Lambda, \psi(q_\Lambda)).$$

As a result, using that $(q_h^*, \psi_h(q_h^*))$ is an optimal pair to Problem (3.36), we get that

$$\beta |D(q_h^*)|(\Omega) \leq -J_0(q_h^*, \psi_h(q_h^*)) + J(q_\Lambda, \psi_h(q_\Lambda)) \leq -m + J_0(q_\Lambda, \psi_h(q_\Lambda)),$$

and thus the sequence $(q_h^*)_h \subset U_{\Lambda,h} \subset U_\Lambda$ is bounded in $BV(\Omega)$ uniformly with respect to h. We denote by $q^* \in U_\Lambda$ its weak* limit and Theorem 3.4.3 then shows that $\psi_h(q_h^*) \to \psi(q^*)$ strongly in $H^1(\Omega)$. The lower semi-continuity of J ensures that

$$J(q^*, \psi(q^*)) = \widetilde{J}(q^*) \leq \liminf_{h\to 0} \widetilde{J}(q_h^*) = \liminf_{h\to 0} J(q_h^*, \psi_h(q_h^*)).$$

Now, let $q \in U_\Lambda$, using the density of smooth function in BV, we get that there exists a sequence $q_h \in U_{\Lambda,h}$ such that $\|q_h - q^*\|_{BV(\Omega)} \to 0$ (see also [9, p.10, Remark 4.2]). From Theorem 3.4.3, we get $\psi_h(q_h) \to \psi(q)$ strongly in $H^1(\Omega)$ and the continuity of J ensure that $\widetilde{J}(q_h) \to \widetilde{J}(q)$. Since $\widetilde{J}(q_h^*) \leq \widetilde{J}(q_h)$ for all $q_h \in U_{\Lambda,h}$, we get by passing to the inf-limit that

$$\widetilde{J}(q^*) \leq \liminf_{h\to 0} \widetilde{J}(q_h^*) \leq \liminf_{h\to 0} \widetilde{J}(q_h) = \widetilde{J}(q), \ \forall q \in U_\Lambda,$$

and the proof is therefore finished. □

When

$$U = U_{\Lambda,\kappa}, \ U_h = U_{\Lambda,\kappa} \cap \mathcal{K}_{h,0},$$

the set of discrete control is composed of piecewise constant function on $\mathcal{T}_h$ that satisfy

$$\forall q_h \in U_h, \ \|q_h\|_{BV(\Omega)} \leq 2\max(\Lambda, \kappa, |\alpha - 1|).$$

We can compute explicitly the previous norm by integrating by parts the total variation (see e.g. [9, p.7, Lemma 4.1]). This reads

$$\forall q_h \in U_h,\ |Dq_h|(\Omega) = \sum_{F\in\mathcal{F}^i} |F||[q_h]|_F|,$$

where $\mathcal{F}^i$ is the set of interior faces and $|[q_h]|_F = |q_h|_{T_1} - q_h|_{T_2}|$ is the jump of q_h on the interior face $F = \partial T_1 \cap \partial T_2$. Note then that any $q_h \in U_h$ can only have either a finite number of discontinuity or jumps that are not too large.

Theorem 3.4.5. *Assume that $\beta = 0$ and $(A2) - (A3)$ from Theorem 3.2.7 hold and that the cost function $J : (q,\psi) \in U_\Lambda \times H^1(\Omega) \mapsto J(q,\psi) \in \mathbb{R}$ is continuous with respect to the* (weak*, weak) *topology of $BV(\Omega) \times H^1(\Omega)$. Let $(q_h^*, \psi_h(q_h^*)) \in U_h \times \mathcal{V}_h$ be an optimal pair of (3.35). Then the sequence $(q_h^*)_h \subset U_{\Lambda,\kappa}$ is bounded and there exists $q^* \in U_{\Lambda,\kappa}$ such that $q_h^* \rightharpoonup q^*$ weakly* in $BV(\Omega)$, $\psi(q_h^*) \to \psi(q^*)$ strongly in $H^1(\Omega)$ and*

$$\widetilde{J}(q^*) \leq \widetilde{J}(q),\ \forall q \in U_\Lambda.$$

Hence any accumulation point of $(q_h^*, \psi_h(q_h^*))$ is an optimal pair for Problem (3.24).

Proof. Since $(q_h^*)_h$ belong to U_h, it satisfies $\|q_h\|_{BV(\Omega)} \leq 2\max(\Lambda, \kappa, |\alpha - 1|)$ and is thus bounded uniformly with respect to h. We denote by $q^* \in U_{\Lambda,\kappa}$ its weak* limit. Theorem 3.4.4 then shows that $\psi_h(q_h^*)$ converges strongly in $H^1(\Omega)$ toward $\psi(q^*)$.

Now, let $q \in U_{\Lambda,\kappa}$, using the density of smooth function in BV, one gets that there exists a sequence $q_h \in U_h$ such that $q_h \rightharpoonup q$ weak* in $BV(\Omega)$ (see also [9, Introduction]). From Theorem 3.4.3, we get $\psi_h(q_h) \to \psi(q)$ strongly in $H^1(\Omega)$ and the continuity of J ensure that $\widetilde{J}(q_h) \to \widetilde{J}(q)$. The proof can then be done as in Theorem 3.4.4. □

3.5 Numerical experiments

In this section, we tackle numerically the optimization problem (3.24), when it is constrained to the total amplitude ψ_{tot} described by (3.18). We focus on two examples: a *damping problem*, where the computed bathymetry optimally reduces the magnitude of the incoming waves; and an *inverse problem*, in which we recover the bathymetry by minimizing the mismatch between the observed and predicted magnitude of the waves.

In what follows, we consider an incident plane wave $\psi_0(x) = \mathrm{e}^{\mathrm{i}k_0 x\cdot\vec{d}}$ propagating in the direction $\vec{d} = (0\ \ 1)^\top$, with

$$k_0 = \frac{\omega_0}{\sqrt{g z_0}},\ \omega_0 = \frac{2\pi}{T_0},\ T_0 = 20,\ g = 9.81,\ z_0 = 3.$$

For the space domain, we set $\Omega = [0, L]^2$, where $L = \frac{10\pi}{k_0}$. We also impose a L^∞-constraint on the variable q, namely that $q \geq -0.9$.

3.5.1 Numerical methods

We discretize the space domain by using a structured triangular mesh of 8192 elements, that is a space step of $\Delta x = \Delta y = 8.476472$.

For the discretization of ψ_{sc}, we use a $\mathbb{P}^1$-finite element method. The optimized parameter q is discretized through a $\mathbb{P}^0$-finite element method. Hence, on each triangle, the approximation of ψ_{sc} is determined by three nodal values, located at the edges of the triangle, and the approximation of q is determined by one nodal value, placed at the center of gravity of the triangle.

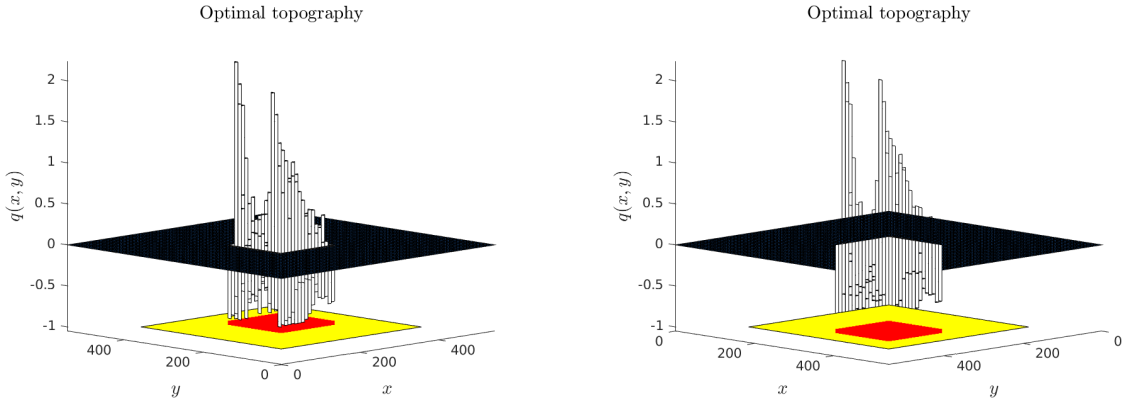
On the other hand, we perform the optimization through a subspace trust-region method, based on the interior-reflective Newton method described in [23] and [22]. Each iteration involves the solving of a linear system using the method of preconditioned conjugate gradients, for which we supply the Hessian multiply function. The computations are achieved with `MATLAB` (version 9.4.0.813654 (R2018a)).

3.5.2 Example 1: a wave damping problem

We first consider the minimization of the cost functional

$$J(q, \psi_{tot}) = \frac{\omega_0^2}{2} \int_{\Omega_0} |\psi_{tot}(x, y)|^2 dxdy,$$

where $\Omega_0 = [\frac{L}{6}, \frac{5L}{6}]^2$ is the domain where the waves are to be damped. The bathymetry is only optimized on a subset $\Omega_q = [\frac{L}{4}, \frac{3L}{4}]^2 \subset \Omega_0$.

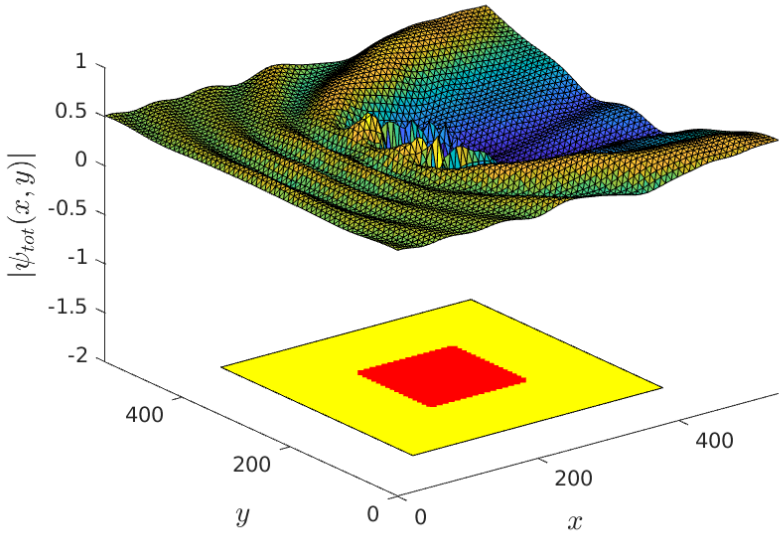


(a) View from above.

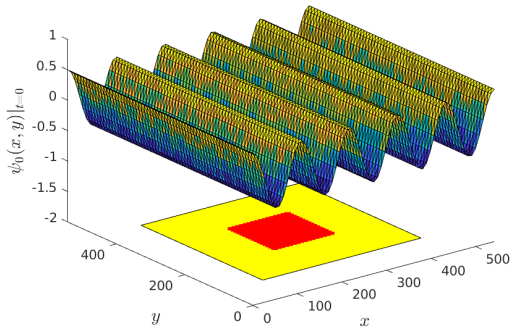
(b) View from below.

Figure 3.1: Optimal topography for a wave damping problem. The yellow part represents Ω_0 and the red part corresponds to the nodal points associated with q.

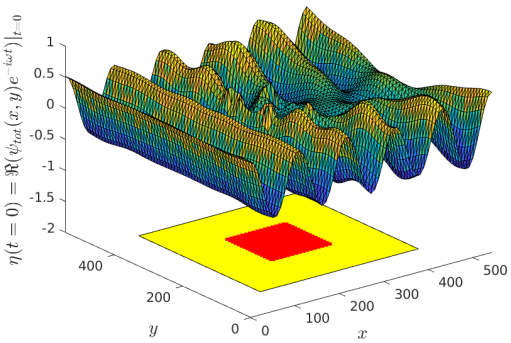
We observe in Figure 3.1 that the optimal bathymetry we obtain is highly oscillating. In our experiments, this oscillation remained at every level of space discretization we have tested. This could be related to the fact that in all our results, $q \in BV(\Omega)$. Note also that the damping is more efficient over Ω_q. This fact is coherent with the results of the next experiment.



(a) Norm of the numerical solution.

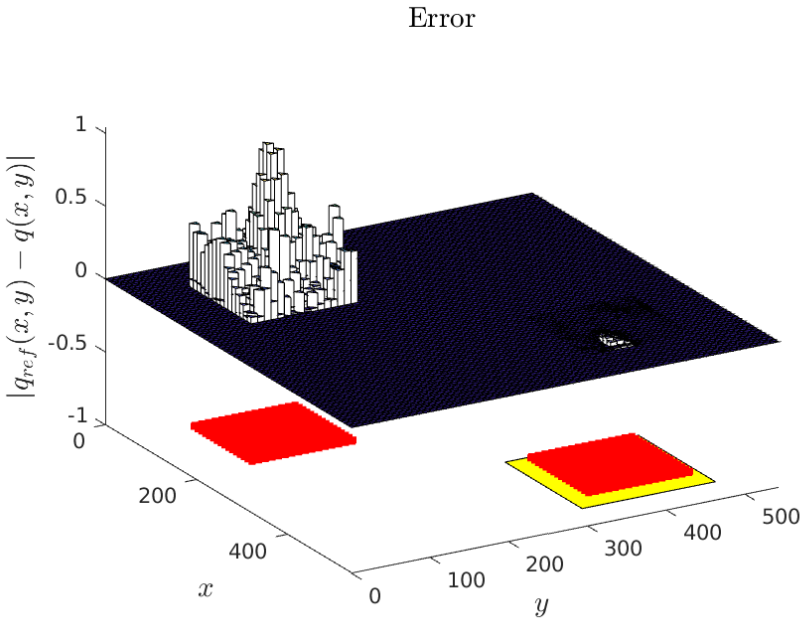


(b) Real part of the incident wave.

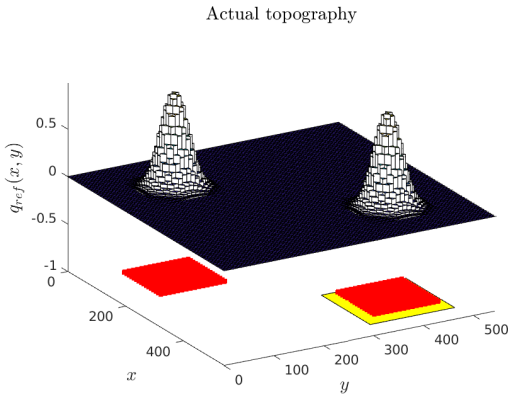


(c) Real part of the numerical solution.

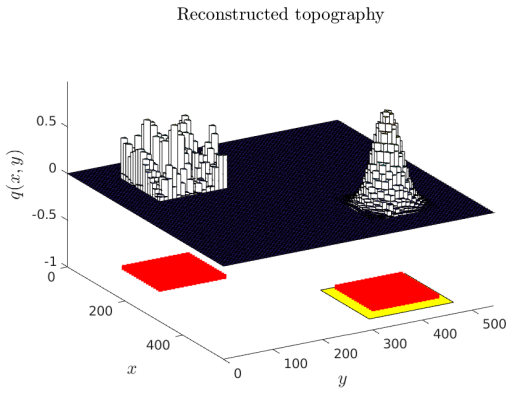
Figure 3.2: Numerical solution of a wave damping problem. The yellow part represents Ω_0 and the red part corresponds to the nodal points associated with q.



(a) Reconstruction error.



(b) Actual bathymetry.



(c) Reconstructed bathymetry.

Figure 3.3: Detection of a bathymetry from a wavefield. The yellow part represents Ω_0 and the red part corresponds to the nodal points associated with q.

3.5.3 Example 2: an inverse problem

Given the bathymetry

$$q_{ref}(x,y) := \mathrm{e}^{-\tau\left(((x-\frac{L}{4})^2+(y-\frac{L}{4})^2\right)} + \mathrm{e}^{-\tau\left((x-\frac{3L}{4})^2+(y-\frac{3L}{4})^2\right)},$$

where $\tau = 10^{-3}$, we use the previous methodology to reconstruct it on the domain $\Omega_q = [\frac{L}{8}, \frac{3L}{8}]^2 \cup [\frac{5L}{8}, \frac{7L}{8}]^2$, by minimizing the cost functional

$$J(q, \psi_{tot}) = \frac{\omega_0^2}{2} \int_{\Omega_0} |\psi_{tot}(x,y) - \psi_{ref}(x,y)|^2 dxdy,$$

where ψ_{ref} is the amplitude associated with q_{ref} and $\Omega_0 = [\frac{3L}{4} - \delta, \frac{3L}{4} + \delta]^2$, $\delta = \frac{L}{6}$. Note that in this case, Ω_q is not contained in Ω_0.

In Figure 3.3, we observe that the part of the bathymetry that does not belong to the observed domain Ω_0 is not recovered by the procedure. On the contrary, the bathymetry is well reconstructed in the part of the domain corresponding to Ω_0.

3.6 Perspectives

The numerical examples highlight two situations in which our approach could be applied. Concerning batrymetry reconstruction, we believe that our standpoint would help in obtaining more precise estimates from observed free surface data. On the other hand, a damped flow is desirable when designing structures as harbors, since minimizes the damage generated by the waves and improves navigation conditions. Then, such a problem a problem could certainly be studied by our analysis.

Finally, the method presented here uses a monochromatic incident wave, focusing entirely on its amplitude. A natural extension is then consider a polychromatic wave, that can be splitted into a sum of monochromatic waves via Fourier transform. Ideally, this should lead us to a family of optimization problems associated with each frequency (that can be solved in parallel), but several questions have to be addressed in between, regarding first a possible decomposition of the cost functional and then the convergence of the whole procedure. Since we work with the Helmholtz equation, this idea cannot be extended to nonlinear wave propagation models as Saint-Venant or Boussinesq. In this case, we suggest to consider the time-dependent problem, associated with time-periodic boundary conditions.

Chapter 4

Mathematical analysis of the Blade element momentum theory

The aim of this chapter is to determine suitable conditions for both the existence of solutions and the numerical solving of the Blade element momentum theory. We recall first the main features of Glauert's modeling, which is followed by the mathematical analysis of the questions above. The key element in our approach is the reformulation of the original set of equations into a single expression, which allows us to split the model into two parts associated with the blade geometry and fluid-turbine dynamic. Last of all, we study the optimization problem concerning turbine efficiency.

This is a joint work with Jérémy Ledoux (Hydrotube Énergie) and Julien Salomon (ANGE, INRIA Paris), via the Grant ANR HyFloEFlu (Hydroliennes flottantes et énergie, ANR-10-IEED-0006-04).

4.1 The Blade element momentum theory

We present the model proposed by Glauert to describe the interaction between a turbine and a flow. After having introduced the relevant variables, we recall the main steps of the reasoning leading to the governing equations. We then detail the simple and complex versions of the model that we will consider.

4.1.1 Variables

Glauert's theory aims at establishing algebraic relations that represent the interaction between a stream and several rotating blades, named *turbine* in what follows. In order to do this, it couples two models: a macroscopic one that describes the evolution of fluids rings crossing the turbine; and a local model that characterizes the behavior of a planar section of a blade (a *blade element*) for various angles of attack and, in some cases, for various Reynolds numbers.

The stream is supposed to be horizontal, constant in time and incompressible. The latter assumption implies that the stream velocities in the left and right neighborhoods of the turbine are equal. We denote them by U_0 and by $U_{-\infty}$ and $U_{+\infty}$ the upstream and downstream velocities, respectively.

As BEM models do not take into account interactions between blade elements, w.l.o.g. we restrict ourselves to a fixed blade element and a constant rotation speed Ω, i.e. a fixed value of the parameter

$$\lambda := \frac{\Omega r}{U_{-\infty}},$$

where Ω denotes the rotation speed of the blades and r is the distance of the element under consideration to the rotation axis. It should also be noted that in practical cases, the turbine works *at constant* λ: Ω is indeed often controlled through the torque exerted by a generator in such a way that the ratio $\frac{\Omega}{U_{-\infty}}$ is kept constant for various values of $U_{-\infty}$. It follows that the value of λ associated with one element only depends on r. In the sequel, we consequently only use the variable λ to describe the location of a blade element.

Macroscopic variables and unknowns

Glauert's model ultimately consists in linking three variables a, a' and φ associated with the ring of fluid under consideration. Among these three unknowns, the *axial induction factor* a and the *tangential induction factor* a' are defined by

$$a = \frac{U_{-\infty} - U_0}{U_{-\infty}}, \tag{4.1}$$

$$a' = \frac{\omega}{2\Omega}, \tag{4.2}$$

where ω is the rotation speed of the ring of fluid at location of the turbine. The angle φ is the *relative angle deviation* [68, p.120] of the ring so that

$$\tan\varphi = \frac{1-a}{\lambda(1+a')}. \tag{4.3}$$

For the sake of simplicity, and to emphasize their role of unknowns in Glauert's model, we omit in this paper the dependence of a, a', φ (and α in what follows) on λ in the notations.

Local variables

Given a blade profile, the *lift* and *drag coefficients* C_L and C_D are defined through the relations

$$dL = C_L(\alpha)\frac{\rho}{2}U_{rel}^2 c_\lambda dr,$$
$$dD = C_D(\alpha)\frac{\rho}{2}U_{rel}^2 c_\lambda dr,$$

where ρ is the mass density of the fluid, dL and dD are the elementary lift and drag forces applying on a blade element of thickness dr and chord $c_\lambda = c_{\lambda(r)}$, and U_{rel} is

the relative fluid speed (also called *apparent fluid speed*) perceived from this blade element while rotating, that is

$$U_{rel} = \frac{U_0}{\sin\varphi}. \tag{4.4}$$

The parameter α is called *angle of attack* and is defined as the angle between the chord and flow direction. It satisfies the relation

$$\alpha = \varphi - \gamma_\lambda, \tag{4.5}$$

where $-\pi/2 \leq \gamma_\lambda \leq \pi/2$ is the *twist* (also called *local pitch*) angle of the blade. The notations associated with a blade element are summarized in Figure 4.1.

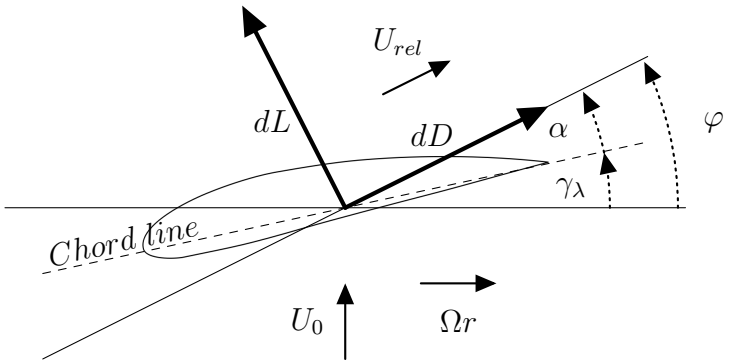


Figure 4.1: Blade element profile and associated angles, velocities and forces

The coefficients C_L and C_D correspond to the ratio of lift and drag forces with respect to the dynamic force, which is associated with the observed kinetic energy. They are assumed to be known, both depending on α and in some cases of the Reynolds number. In the following, we do not take into account their dependence on the latter, however, all the results established in this chapter can be extended without difficulty to that framework. On the contrary to α, the Reynolds number is known since it depends on α as well as $U_{-\infty}$, r, Ω, c_λ (see [87, p.374]), variables that can be controlled e.g. in a wind tunnel.

Thought varying from one profile to another, the variations of C_L and C_D with respect to α can be described qualitatively in a general way. The coefficient C_L usually increases linearly with respect to α up to a given critical angle α_s, with $0 < \alpha \leq \pi/2$, where the so-called *stall* phenomenon occurs: C_L then decreases rapidly (see, e.g., [15, pp.93-94]), causing a sudden loss of lift, so that angles of attack larger that α_s are not desirable. Since C_D is associated with a drag force, it is always positive and defined for all angles in concrete cases. Its usual behavior consists in slightly increasing with α up to $\alpha = \alpha_s$ where it then becomes very large. As a consequence, the condition $\varphi < \alpha_s + \gamma_\lambda$ is always required in the blade design phase.

We summarize the properties of C_L and C_D required for our study in the following assumption.

Assumption 4.1.1. *In what follows, we assume that C_L is well-defined and continuous on an interval*

$$I_\beta := [-\beta, \beta]$$

for some $\beta \in [0, \alpha_s)$ and positive on $I_\beta \cap \mathbb{R}^+$. The coefficient C_D is well-defined and positive on $\mathbb{R}$.

It should be noted that as soon as we set the blade profile, that is, C_L and C_D, the chord c_λ and twist γ_λ are the main design parameters. Their optimization will be discussed in Section 4.4.

4.1.2 Glauert's modeling

For the sake of completeness, we now shortly recall here the reasoning proposed by Glauert to model the interaction between a turbine and a stream. We refer to [20, Chapter 3] for a more extended presentation of the theory. We denote by dT the infinitesimal thrust and dQ the infinitesimal torque that apply on the blade element under consideration.

Macroscopic approach

The first part of the model is related to the *Momentum Theory*, dealing with the macroscopic evolution of the ring of fluid. It aims at expressing dT and dQ in terms of a, a' and φ. Denoting by p_- and p_+ the fluid pressure left and right neighborhoods of the turbine, we apply twice Bernouilli's relation between $-\infty$ and 0^- and between 0^+ and $+\infty$ to get

$$p_- - p_+ = \frac{1}{2}\rho(U_{-\infty}^2 - U_{+\infty}^2).$$

Considering then the rate of change of momentum through the turbine, we obtain a second expression for the variation in the pressure, namely

$$p_- - p_+ = \rho(U_{-\infty} - U_{+\infty})U_0.$$

Combining the two latter equations and using (4.1), we get

$$U_{+\infty} = (1 - 2a)U_{-\infty}.$$

Since $dT = (p_- - p_+)\pi r dr$ and $dQ = \omega\rho U_0 2\pi r^3 dr$, we finally obtain

$$dT = 4a(1-a)U_{-\infty}^2 \rho\pi r dr, \tag{4.6}$$
$$dQ = 4a'(1-a)\lambda U_{-\infty}^2 \rho\pi r^2 dr. \tag{4.7}$$

The quantity

$$C_T = \frac{dT}{\frac{1}{2}U_{-\infty}^2 \rho 2\pi r dr}, \tag{4.8}$$

is often called *local thrust coefficient* [91].

Local expressions

Another set of equations can be obtained through the Blade Element Theory, where local expressions for infinitesimal thrust and torque are considered. The approach can be summarized as follows: by (4.5), we have

$$dT = \frac{B}{2} U_{rel}^2 \left(C_L(\varphi - \gamma_\lambda)\cos\varphi + C_D(\varphi - \gamma_\lambda)\sin\varphi\right) \rho c_\lambda dr,$$
$$dQ = \frac{B}{2} U_{rel}^2 \left(C_L(\varphi - \gamma_\lambda)\sin\varphi - C_D(\varphi - \gamma_\lambda)\cos\varphi\right) \rho c_\lambda r dr,$$

where B is the number of blades of the turbine under consideration. These equations can then be combined with (4.4) to give

$$dT = \sigma_\lambda \frac{(1-a)^2}{\sin^2\varphi} \left(C_L(\varphi - \gamma_\lambda)\cos\varphi + C_D(\varphi - \gamma_\lambda)\sin\varphi\right) U_{-\infty}^2 \rho\pi r dr, \tag{4.9}$$
$$dQ = \sigma_\lambda \frac{(1-a)^2}{\sin^2\varphi} \left(C_L(\varphi - \gamma_\lambda)\sin\varphi - C_D(\varphi - \gamma_\lambda)\cos\varphi\right) U_{-\infty}^2 \rho\pi r^2 dr, \tag{4.10}$$

with $\sigma_\lambda = \frac{Bc_\lambda}{2\pi r}$.

Glauert's relations

To close the system of equations, Glauert combined the previous results. More precisely, equating (4.6) and (4.7) with (4.9) and (4.10) respectively, and dividing both resulting equations by $4(1-a)^2$ gives

$$\frac{a}{1-a} = \frac{\sigma_\lambda}{4\sin^2\varphi} \left(C_L(\varphi - \gamma_\lambda)\cos\varphi + C_D(\varphi - \gamma_\lambda)\sin\varphi\right), \tag{4.11}$$
$$\frac{a'}{1-a} = \frac{\sigma_\lambda}{4\lambda\sin^2\varphi} \left(C_L(\varphi - \gamma_\lambda)\sin\varphi - C_D(\varphi - \gamma_\lambda)\cos\varphi\right). \tag{4.12}$$

The system obtained by assembling (4.3), (4.11) and (4.12) is the basis of Glauert's Blade element momentum theory.

4.1.3 Simplified model

In the monographs devoted to aerodynamics of wind turbines, the contribution of C_D is ocasionally set to zero. This point is discussed in [86, p.135], where it is particular stated that *"Since the drag force does not contribute to the induced velocity physically, C_D is usually omitted when calculating induced velocities."*

In the same way, Manwell mentions in [68, p.125]: *"In the calculation of induction factors,[...] accepted practice is to set C_D equal to zero [...]. For airfoils with low drag coefficients, this simplification introduces negligible errors."*.

This assumption is actually justified in many cases, since the procedures used to design profiles aim at minimizing their drag. Moreover, as explained in Section 4.4.1, the usual blade design procedure starts by selecting a twist angle γ_λ that minimizes the ratio $\frac{C_D}{C_L}$.

We also consider this case, referred as *simplified model* in the following. It corresponds to the three equations:

$$\tan\varphi = \frac{1-a}{\lambda(1+a')}, \tag{4.13}$$

$$\frac{a}{1-a} = \frac{1}{\sin^2\varphi}\mu_L(\varphi)\cos\varphi, \tag{4.14}$$

$$\frac{a'}{1-a} = \frac{1}{\lambda\sin\varphi}\mu_L(\varphi), \tag{4.15}$$

where we have introduced the dimensionless function $\mu_L(\varphi) := \frac{\sigma_\lambda}{4}C_L(\varphi-\gamma_\lambda)$, that is defined on

$$I_{\beta,\gamma_\lambda} := [-\beta+\gamma_\lambda, \beta+\gamma_\lambda] \tag{4.16}$$

by virtue of Assumption 4.1.1.

4.1.4 Corrected model

To get closer to the experimental results, many modifications of the model defined by (4.3), (4.11) and (4.12) have been introduced, see e.g. [86, Section 3]. Hereafter, we present three important corrections, namely non-zero drag coefficient C_D, tip loss correction and a specific treatment of a for cases where its values become large. The first and last will modify significantly the analysis developed for the simplified Glauert's model.

Weak drag

In the corrected model that we will consider, we assume that C_D is strictly positive. However, in our analysis, we will assume small values of this parameter and/or a slow increasing before the occurrence of the stall phenomenon.

Tip loss correction

The independence of the infinitesimal rings of fluid, that is assumed in the model of Glauert is valid for rotors with *infinite* many blades. In real situations, a modification of the flow at the tip of a blade has to be included to take into account that the circulation of the fluid around the blade must go down (exponentially) to zero. In this way, given a number of blades B and a radius R of the considered turbine, Glauert

(see [43, p.268]) has used the Prandtl tip function F_λ [77]:

$$F_\lambda(\varphi) := \frac{2}{\pi}\cos^{-1}\left(\exp(-\frac{\frac{B}{2}(1-\frac{r}{R})}{\frac{r}{R}\sin\varphi})\right),$$

as a new factor in Equations (4.6) and (4.7), which give rise to

$$dT = 4a(1-a)F_\lambda(\varphi)U_{-\infty}^2\rho\pi r dr, \tag{4.17}$$
$$dQ = 4a'(1-a)F_\lambda(\varphi)U_{-\infty}\rho\pi r^3\Omega dr. \tag{4.18}$$

Further models of tip loss corrections have been introduced in between. We refer to [84] and [15, Chapter 13], for reviews of these models.

Correction for high values of a

For induction factors a larger than about 0.4, a turbulent wake state generally occurs and it is broadly consider that momentum theory does not apply [87, p.297]. This fact was already noted by Glauert [42], who proposed, for large values of a, to modify the thrust expression (4.8) to fit with experimental observations. Subsequently, many other expressions have been proposed to improve this fitting, see [15, Section 10.2.2].

All these corrections lead to modify the infinitesimal thrust dT and consequently the term $a(1-a)$ in Definition (4.17), that becomes

$$dT = 4\chi(a,a_c)F_\lambda(\varphi)U_{-\infty}^2\rho\pi r dr. \tag{4.19}$$

In the literature, the function $\chi(a,a_c)$ is in most cases of the form

$$\chi(a,a_c) = a(1-a) + \psi\left((a-a_c)_+\right), \tag{4.20}$$

where $(a-a_c)_+ := \max(0, a-a_c)$ and ψ is a given function defined on $\mathbb{R}^+$. Various choices of corrections are presented via the function ψ in Table 4.1. Note that Glauert's empirical correction leads to a discontinuity at $a = a_c$ when $F_\lambda(\varphi) \neq 1$ (see [15, p.195]). Buhl proposed in [19] a slight modification to fix this issue.

Corrected system

We now repeat the reasoning used to obtain Equations (4.13–4.15), that is, we equalize (4.19) and (4.18) respectively with (4.9) and (4.10). This gives, using (4.20) and simplifying:

$$\tan\varphi = \frac{1-a}{\lambda(1+a')}, \tag{4.21}$$
$$\frac{a}{1-a} = \frac{1}{\sin^2\varphi}(\mu_L^c(\varphi)\cos\varphi + \mu_D^c(\varphi)\sin\varphi) - \frac{\psi\left((a-a_c)_+\right)}{(1-a)^2}, \tag{4.22}$$
$$\frac{a'}{1-a} = \frac{1}{\lambda\sin^2\varphi}(\mu_L^c(\varphi)\sin\varphi - \mu_D^c(\varphi)\cos\varphi), \tag{4.23}$$

Order	Authors	a_c	$\psi((a-a_c)_+)$
3	Glauert [43]	1/3	$\frac{(a-a_c)_+}{4}\left(\frac{(a-a_c)_+^2}{a_c}+2(a-a_c)_++a_c\right)$
2	Glauert empirical [50, p.25] [68, p.103]	2/5	$a_c(1-a_c)$ $+\frac{(a-a_c)_+[F_\lambda(\varphi)\,((a-a_c)_++2a_c)-0.286]}{2.5708}F_\lambda(\varphi)$
2	Buhl [19]	2/5	$\frac{1}{2F_\lambda(\varphi)}\left(\frac{(a-a_c)_+}{1-a_c}\right)^2$
1	Wilson et al., Spera [87, p.302]	1/3	$(a-a_c)_+^2$

Table 4.1: Various corrections proposed in the literature. The order corresponds to the degree of a (as a polynomial) in $\chi(a,a_c)$

where we have introduced the dimensionless functions

$$\mu_L^c(\varphi) := \frac{\sigma_\lambda}{4F_\lambda(\varphi)}C_L(\varphi-\gamma_\lambda), \quad \mu_D^c(\varphi) := \frac{\sigma_\lambda}{4F_\lambda(\varphi)}C_D(\varphi-\gamma_\lambda), \tag{4.24}$$

defined respectively on I_{β,γ_λ} and $\mathbb{R}$. Since in usual applications $a \leq 1$, the corrected model coincides with the simplified model when $F_\lambda(\varphi)=1$, $a_c=1$ and $C_D=0$.

4.2 Analysis of Glauert's model and existence of solution

In this section, we reduce each of the two previous versions of Glauert's model to a scalar equation, that explicitly separates local and macroscopic variables (up to one term for the corrected model). This lead us to formulate assumptions that only concern the characteristics of the turbine, that is, on the functions $\mu_L, \mu_D, \mu_L^c, \mu_D^c$, or equivalently, in C_L and C_D.

We define the angle $\theta_\lambda \in (0, \frac{\pi}{2})$ by

$$\tan\theta_\lambda := \frac{1}{\lambda}. \tag{4.25}$$

and

$$\begin{aligned} I &:= I_{\beta,\gamma_\lambda} \cap \left(-\frac{\pi}{2}+\theta_\lambda, \frac{\pi}{2}+\theta_\lambda\right), \\ I^+ &:= I \cap (0,\theta_\lambda]. \end{aligned} \tag{4.26}$$

4.2.1 Simplified model

In this simple setting, a reformulation of Equations (4.13–4.15) can be obtained after a short algebraic manipulation.

Theorem 4.2.1. *Suppose that Assumption 4.1.1 holds and that* $(\varphi, a, a') \in I - \{0, \frac{\pi}{2}\} \times \mathbb{R} - \{1\} \times \mathbb{R} - \{-1\}$ *satisfies Eqs (4.13–4.15). Then* φ *satisfies*

$$\mu_L(\varphi) = \mu_G(\varphi), \tag{4.27}$$

where

$$\mu_G(\varphi) := \sin\varphi \tan(\theta_\lambda - \varphi).$$

Reciprocally, suppose that $\varphi \in I - \{0, \frac{\pi}{2}\}$ *satisfies Eq. (4.27) and define* a *and* a' *as the corresponding solutions of Eqs. (4.14) and (4.15), respectively. Then* $(\varphi, a, a') \in I - \{0, \frac{\pi}{2}\} \times \mathbb{R} - \{1\} \times \mathbb{R} - \{-1\}$ *satisfies Eqs. (4.13–4.15).*

Equation (4.27) appears – up to a factor – in [68, p.128, Equation (3.85a)]. We see that the intervals I_{β,γ_λ} and $(-\frac{\pi}{2} + \theta_\lambda, \frac{\pi}{2} + \theta_\lambda)$ play similar roles in the local and macroscopic descriptions, respectively, as they both correspond to domains of definition. In the same way, $I_{\beta,\gamma_\lambda} \cap \mathbb{R}^+$ and $(0, \theta_\lambda]$, whose intersection is I^+, correspond to angles associated with positive lift in the two descriptions.

We have excluded the angles $\varphi = 0$ and $\varphi = \frac{\pi}{2}$ for the sole reason that Equations (4.13–4.15) are not defined correctly for these values. Actually, $\varphi = 0$ is naturally associated with the case $a = 1$. On the contrary, the value $\varphi = \frac{\pi}{2}$ (that belongs to I if $\beta + \gamma_\lambda > \frac{\pi}{2}$) is neither a solution of Equations (4.13–4.15) nor of (4.27): it leads indeed to $a' = -1$, $a = 0$ and $-\lambda = \mu_L(\frac{\pi}{2})$ which corresponds to a negative lift, contradicting Assumption 4.1.1. Consequently, both values can be included in the interval of admissible angles without introducing a spurious solution. In addition, these angles do not pose any particular problems concerning the definition of μ_G, so that the formulation (4.27) will be considered on I in the rest of this chapter. Note also that $\varphi = \theta_\lambda \pm \frac{\pi}{2}$ are neither solutions of Equations (4.13–4.15) nor of (4.27), since they lead to an absurd.

Proof. Suppose that $(\varphi, a, a') \in I \times \mathbb{R} - \{1\} \times \mathbb{R} - \{-1\}$ satisfies Equations (4.21–4.23). We have to prove that φ satisfies (4.27). Eliminating a and a' in (4.13), using (4.14) and (4.15), we get

$$\begin{aligned}\tan^{-1}\varphi = \lambda\frac{1+a'}{1-a} &= \lambda(1 + \frac{a}{1-a}) + \lambda\frac{a'}{1-a} \\ &= \lambda\left(1 + \frac{\cos\varphi}{\sin^2\varphi}\mu_L(\varphi)\right) + \frac{1}{\sin\varphi}\mu_L(\varphi).\end{aligned}$$

so that (4.27) follows from Definition (4.25) of θ_λ. Repeating these steps backward ends the proof of the equivalence. □

This result shows that Glauert's model – here in its simplified version – essentially boils down to an only scalar equation: indeed, suppose that φ satisfies Equation (4.27), then a and a' can be post-computed thanks to Equations (4.14–4.15). These quantities are consequently a by-product of the determination of φ.

An important property of Equation (4.27) is that its left-hand side corresponds to the local description of the problem, whereas the right-hand side is rather related to the macroscopic modeling introduced by Glauert's theory, where μ_G appears to be a universal function in fluid-turbine dynamics. Another way to formulate this fact consists in associating the left-hand side with 2D experimental data (through C_L and C_D) and design variables (as γ_λ or σ_λ), and the right-hand side with a 3D model and the parameter θ_λ, that does not depend on the blade geometry. This splitting is the heart of Glauert's model.

The formulation given in Theorem 4.2.1 gives rise to many criteria to ensure existence of solution of (4.27): the existence indeed holds as soon as the graphs of μ_G and μ_L intersect. As an illustration, we only give a simple condition in the case of symmetric profiles.

Corollary 4.2.2. *In addition to Assumption 4.1.1, suppose that the profile under consideration is symmetric with $\gamma_\lambda > 0$, and that*

$$\mu_G(\max I) \leq \mu_L(\max I), \tag{4.28}$$

where $\max I = \min\{\theta_\lambda, \beta + \gamma_\lambda\}$. Then (4.27) admits a solution in $[\gamma_\lambda, \max I]$ corresponding to a positive lift. Moreover, if $\max I = \theta_\lambda$, i.e. $\theta_\lambda \leq \beta + \gamma_\lambda$, then (4.28) is automatically satisfied.

Proof. The set I is non-empty as soon as

$$\theta_\lambda - \gamma_\lambda - \frac{\pi}{2} \leq \beta.$$

Since $\gamma_\lambda > 0$ and $\beta > 0$, the equation above holds, so that $\max I$ is well defined. As we consider a symmetric profile, we have $\mu_L(\gamma_\lambda) = C_L(0) = 0$ whereas $\mu_G(\gamma_\lambda) > 0$. Due to the continuity assumption on μ_L and (4.28), the existence of solution of (4.27) in $[\gamma_\lambda, \max I]$ follows from the Intermediate Value Theorem. The positivity of μ_G on the interval $[\gamma_\lambda, \max I]$ implies that the resulting lift is positive.

If $\max I = \theta_\lambda$, then $\mu_G(0) = 0$ and μ_L is positive on $[\gamma_\lambda, \max I]$, the last assertion follows. □

In the case where μ_L is supplementary assumed to be increasing on $[\gamma_\lambda, \beta + \gamma_\lambda]$, then, the solution defined in Theorem 4.2.2 is unique.

4.2.2 Corrected model

We now consider the corrected model defined by Equations (4.21–4.23), for a given value $a_c \in (0,1)$. The algebraic manipulations performed in the previous section to get Lemma 4.2.1 cannot be pushed as far as in the simplified model and lead to expressions still containing the unknown a. Hence, before stating a reformulation of this model and an existence result, we need to clarify the dependence of a on the variable φ. Again, we express our assumptions in terms of μ_L^c and μ_D^c, but the translation in terms of C_L, C_D, σ_λ and $F_\lambda(\varphi)$ is straightforward.

In all this section, we suppose that $0 \in I$, i.e. $|\gamma_\lambda| \leq \beta$, which means in particular that

$$I^+ = (0, \min\{\theta_\lambda, \beta + \gamma_\lambda\}], \tag{4.29}$$

Lemma 4.2.3. *Suppose that Assumption 4.1.1 holds and define, for $\varphi \in I^+$*

$$g(\varphi) := \tan^{-1}\varphi \tan(\theta_\lambda - \varphi) + \frac{\mu_D^c(\varphi)}{\sin\varphi}\left(1 + \tan^{-1}\varphi\tan(\theta_\lambda - \varphi)\right). \tag{4.30}$$

Let ψ be one of the functions given in Table 4.1, with $F_\lambda(\varphi) = 1$ in the case of Glauert empirical correction. Then, the expression

$$\frac{a}{1-a} + \left(1 - \frac{\cos\theta_\lambda\cos\varphi}{\cos(\theta_\lambda - \varphi)}\right)\frac{\psi\left((a-a_c)_+\right)}{(1-a)^2} = g(\varphi) \tag{4.31}$$

defines a continuous decreasing mapping $\tau : \varphi \in I^+ \mapsto a \in [0,1)$.

Moreover, if g is decreasing and μ_D^c differentiable, then τ is decreasing and differentiable in all points with a possible exception of one point φ_c.

As a by-product of the properties of C_D stated in Assumption 4.1.1, the function μ_D^c is always positive and defined for all angles in concrete cases so that g is well defined on I^+. Note also that the only obstruction for g to be decreasing would come from the term C_D which is often increasing in a neighborhood of 0. But for usual profiles, its variations are negligible when compared to the other (decreasing) terms in (4.30).

Proof. To simplify the notations, let us rewrite Equation (4.31) under the form

$$u(a) + v(\varphi)w(a) = g(\varphi), \tag{4.32}$$

with

$$u(a) := \frac{a}{1-a},\ v(\varphi) := 1 - \frac{\cos\theta_\lambda\cos\varphi}{\cos(\theta_\lambda - \varphi)},\ w(a) := \frac{\psi\left((a-a_c)_+\right)}{(1-a)^2}.$$

Let us first consider the left-hand side of (4.32). We see that u is positive and increasing on $[0,1)$ as well as w for any function ψ given in Table 4.1 (with $F_\lambda(\varphi) = 1$

in the case of Glauert empirical correction). In the same way, v is positive and increasing on $(0, \theta_\lambda]$. Fixing now $\varphi \in (0, \theta_\lambda]$, it is fairly easy to see that the mapping $a \in [0, 1) \mapsto u(a) + v(\varphi)w(a)$ is continuous, strictly increasing (actually strictly convex), strictly positive and goes from 0 to $+\infty$. Since g is assumed to be positive on I^+, there exists an only a in $[0, 1)$ such that (4.32) holds. Hence the existence of the mapping τ.

Suppose now that g is decreasing and μ_D^c differentiable. Except w in the point $a = a_c$, all the functions involved in Equation (4.32) are differentiable. Consider $\varphi \in I^+$, such that $\tau(\varphi) \neq a_c$. The functions u and w are differentiable in $a = \tau(\varphi) \in (0, 1)$ and $u'(a) + v(\varphi)w'(a) \neq 0$ so that we can differentiate Equation (4.32) with respect to φ. We get

$$\tau'(\varphi) = \frac{g'(\varphi) - v'(\varphi)w\left(\tau(\varphi)\right)}{u'\left(\tau(\varphi)\right) + v(\varphi)w'\left(\tau(\varphi)\right)}.$$

Combining the fact that g is decreasing with the above properties of v, w, u and their derivatives implies that $\tau'(\varphi) \leq 0$. As a consequence, the mapping τ is either differentiable on the whole interval I^+, or differentiable on a set of the form $I^+ - \{\varphi_c\}$ where φ_c is the only value in I^+ such that $\tau(\varphi_c) = a_c$ The result follows. □

Remark 4.2.4. *The quantity $a = \tau(\varphi)$ can generally be computed explicitly provided that the function ψ is specified analytically, as e.g. in Table 4.1. In the last case, the computation consists in solving a low order polynomial equation in a.*

We can now state a result similar to Theorem 4.2.1 in the case of the corrected model.

Theorem 4.2.5. *Let the assumptions of Lemma 4.2.3 hold so that τ is defined on I^+ and ψ be one of the functions given in Table 4.1, with $F_\lambda(\varphi) = 1$ in the case of Glauert empirical correction. Suppose also that $(\varphi, a, a') \in I^+ - \{\frac{\pi}{2}\} \times \mathbb{R} - \{1\} \times \mathbb{R}$ satisfies Equations (4.21–4.23). Then φ satisfies*

$$\mu_L^c(\varphi) - \tan(\theta_\lambda - \varphi)\mu_D^c(\varphi) = \mu_G^c(\varphi), \tag{4.33}$$

where

$$\mu_G^c(\varphi) := \mu_G(\varphi) + \frac{\cos\theta_\lambda \sin^2\varphi}{\cos(\theta_\lambda - \varphi)} \frac{\psi\left((\tau(\varphi) - a_c)_+\right)}{(1 - \tau(\varphi))^2}. \tag{4.34}$$

Reciprocally, suppose that $\varphi \in I^+ - \{\frac{\pi}{2}\}$ satisfies (4.33) and define a and a' by

$$a = \tau(\varphi), \tag{4.35}$$

$$a' = \frac{1 - \tau(\varphi)}{\lambda \sin^2\varphi}(\mu_L^c \sin\varphi - \mu_D^c \cos\varphi). \tag{4.36}$$

Then $(\varphi, a, a') \in I^+ \in I^+ - \{\frac{\pi}{2}\} \times \mathbb{R} - \{1\} \times \mathbb{R}$ satisfies Equations (4.21–4.23).

As it was the case with the simplified model, the value $\varphi = \frac{\pi}{2}$ is excluded only for the technical reason that (4.21) is not defined for this angle.

Proof. Let $(\varphi, a, a') \in I^+ - \{\frac{\pi}{2}\} \times [0,1) \times \mathbb{R}^+$ satisfiying Equations (4.21–4.23). Because of (4.21), we get:

$$\tan^{-1}\varphi = \lambda \frac{1+a'}{1-a} = \lambda(1 + \frac{a}{1-a}) + \lambda \frac{a'}{1-a}.$$

In the latter, the terms $\frac{a}{1-a}$ and $\lambda\frac{a'}{1-a}$ can be eliminated thanks to (4.22) and (4.23), respectively. After some algebraic manipulations, we end up with a corrected version of (4.27):

$$\mu_L^c = (\sin\varphi + \mu_D^c)\tan(\theta_\lambda - \varphi) + \frac{\cos\theta_\lambda \sin^2\varphi}{\cos(\theta_\lambda - \varphi)} \frac{\psi\left((a-a_c)_+\right)}{(1-a)^2},$$

so that (4.33) is satisfied. Using now the last expression to eliminate μ_L^c in (4.22) gives (4.31). Consequently, Lemma 4.2.3 implies that a and φ satisfy (4.35). Finally, (4.36) is a direct consequence of (4.35) and (4.23).

Suppose now that $(\varphi, a, a') \in I^+ - \{\frac{\pi}{2}\} \times [0,1) \times \mathbb{R}^+$ satisfies Equations (4.33–4.36). Replacing $\tau(\varphi)$ by a in (4.36) gives immediately (4.23). Combining (4.35) with the definition of μ_G^c give (4.22). Finally, (4.21) is obtained by introducing a and a' thus defined into (4.33). □

As in the simplified model, Glauert's model boils down to an only scalar equation, with φ as an unknown. However, on the contrary to the simplified model, the formulation (4.33) does not completely decompose the terms into a local part and macroscopic modelling part: much as the left-hand side of Equation (4.33) still relies on local features and experimental 2D data of the problem, its right-hand side now also contains an experimental term, namely μ_D^c through the definition of τ.

Before going further, let us give more details about the behavior of τ in $\varphi = 0$.

Lemma 4.2.6. *Let Assumption 4.1.1 hold so that τ is defined on I^+ and ψ be one of the functions given in Table 4.1, with $F_\lambda(\varphi) = 1$ in the case of Glauert empirical correction. Then*

$$\tau(\varphi) = 1 - \sqrt{\frac{\psi(1-a_c)}{\mu_D^c(0)}}\varphi^{3/2} + o_{\varphi=0}(\varphi^{3/2}).$$

Proof. Let us first prove that $\lim_{\varphi\to 0^+} \tau(\varphi) = 1^-$. From Equation (4.30), we see that $\lim_{\varphi\to 0^+} g(\varphi) = +\infty$. Given $\varphi \in (0, \theta_\lambda]$, we have $a = \tau(\varphi) \in [0,1)$ and

$$1 - \frac{\cos\theta_\lambda \cos\varphi}{\cos(\theta_\lambda - \varphi)} = \frac{\sin\theta_\lambda \sin\varphi}{\cos(\theta_\lambda - \varphi)} \geq 0,$$

so that all the terms of the left-hand side of (4.31) are positive. As a consequence, the only possibility for the sum of these terms to go to $+\infty$ is that $\lim_{\varphi\to 0^+} \tau(\varphi) = 1^-$.

Defining $\nu(\varphi) = 1 - \tau(\varphi)$, we expand (4.31) in a neighborhood of $\varphi = 0^+$ to get

$$\frac{1}{\nu(\varphi)} - 1 + \frac{\varphi(\tan\theta_\lambda \cdot \psi(1-a_c) + o_{\varphi=0}(1))}{\nu(\varphi)^2} = \frac{\mu_D^c(0)\tan\theta_\lambda}{\varphi^2} + o_{\varphi=0}(\frac{1}{\varphi^2}),$$

from which we obtain, after some computations,

$$\frac{\nu^2(\varphi)}{\varphi^3}\left(\frac{\varphi^2}{\nu(\varphi)} - \varphi^2 - (o_{\varphi=0}(1) + \mu_D^c(0)\tan\theta_\lambda)\right) = -\tan\theta_\lambda \cdot \psi(1-a_c) + o_{\varphi=0}(1). \tag{4.37}$$

We consider a sequence $(\varphi_n)_{n\in\mathbb{N}}$ satisfying $\lim_{n\to+\infty}\varphi_n = 0^+$, so that $\lim_{n\to+\infty}\nu(\varphi_n) = 0^+$. Assuming

$$\lim_{n\to+\infty}\frac{\nu^2(\varphi_n)}{\varphi_n^3} = +\infty,$$

we obtain that the sequence $\frac{\varphi_n^2}{\nu(\varphi_n)} = \left(\frac{\varphi_n^3}{\nu^2(\varphi_n)}\right)^{2/3}\nu^{1/3}(\varphi_n)$ goes to zero.

Back to (4.37), we find a contradiction since the left-hand side goes to $+\infty$ whereas the right-hand side is constant. It follows that, up to a subsequence, we can assume that $\lim_{n\to+\infty}\frac{\nu^2(\varphi_n)}{\varphi_n^3} = \ell$ for a certain ℓ. Setting $\varphi = \varphi_n$ in (4.37) and passing to the limit $n \to +\infty$, we obtain that $\ell = \frac{\psi(1-a_c)}{\mu_D^c(0)}$. The result follows. □

Remark 4.2.7. *In the case $C_D(0) = 0$, a similar reasoning gives:*

$$\tau(\varphi) = 1 - \sqrt{\psi(1-a_c)\varphi} + o_{\varphi=0}(\varphi^{1/2}).$$

The quantity $\mu_D^c(0)$ has no specific physical meaning in the applications. We have introduced it as a constant (that can be written explicitly), for simplicity of presentation. The angle $\varphi = 0$ has rather a meaning from the macroscopic point of view, as is it appears when considering μ_G that cancels in 0 and μ_G^c, see (4.39) hereafter.

We are now in the position to give an existence result about the corrected model.

Corollary 4.2.8 (of Theorem 4.2.5)**.** *Suppose that assumptions of Lemma 4.2.6 hold and that*

$$\mu_G^c(\max I^+) \le \mu_L^c(\max I^+) - \tan(\theta_\lambda - \max I^+)\mu_D^c(\max I^+). \tag{4.38}$$

Then Equation (4.33) admits a solution in I^+ corresponding to a positive lift. Moreover if g is decreasing, $\max I^+ = \theta_\lambda$ and $\varphi^c < \theta_\lambda$, then the assumption (4.38) is automatically satisfied.

Proof. Because of Lemma 4.2.6 and Definition (4.34) of μ_G^c, we get

$$\mu_G^c(\varphi) \approx_{\varphi \to 0^+} \frac{\mu_D^c(0)}{\varphi},$$

so that

$$\lim_{\varphi \to 0^+} \mu_G^c(\varphi) = +\infty. \tag{4.39}$$

This implies that there exists a small enough $\varphi_0 > 0$ such that $\mu_G^c(\varphi_0) \geq \mu_L^c(\varphi_0) - \tan(\theta_\lambda - \varphi_0)\mu_D^c(\varphi_0)$. Because of (4.38), the existence of a solution of Equation (4.33) follows from the Intermediate Value Theorem. The positivity of μ_G^c on I^+ implies that the resulting lift is positive.

Suppose now that g is decreasing, $\max I^+ = \theta_\lambda$ and $\varphi^c < \theta_\lambda$. Because of the last assertion of Lemma 4.2.6, the mapping τ is a decreasing function of φ, the correction associated with ψ is not any more active on $[\varphi^c, \theta_\lambda)$. We then have

$$\begin{aligned}\mu_G^c(\max I^+) &= \mu_G(\max I^+) = \mu_G(\theta_\lambda) \leq 0 \\ &\leq \mu_L^c(\max I^+) = \mu_L^c(\max I^+) - \tan(\theta_\lambda - \max I^+)\mu_D^c(\max I^+).\end{aligned}$$

As a consequence, Equation (4.38) holds. □

Unlike the simplified model, no condition on γ_λ or $\mu_L^c(\gamma_\lambda)$ is assumed to get the previous (and following) result, but the alternative (non restrictive) Assumption (4.29) is required. This makes the corrected model much better posed than its simplified version.

Remark 4.2.9. *In the case $\mu_D^c(0) = 0$, a similar reasoning gives:*

$$\mu_G^c(\varphi) \approx_{\varphi \to 0^+} (1 + \tan\theta_\lambda)\varphi.$$

As a consequence, $\mu_G^c(0) = 0$, so that as in the simplified model, we need an assumption about, e.g. $\mu_L^c(\gamma_\lambda)$ to get an existence result similar to Corollary 4.2.2.

4.2.3 Multiple solutions

The results presented on the previous sections can be completed with some additional remarks about multiple solutions. We identify at least three independent situations.

Simplified model

Since $\lim_{\varphi \to \theta_\lambda \pm \pi/2} \mu_G(\varphi) = -\infty$, there shall be two intersections between the graphs of μ_G and μ_L, e.g. in the case where μ_L is affine, $C_L(0) = 0$ and $\gamma_\lambda \in (0, \theta_\lambda]$. In this framework, one of the two roots gives rise to a negative lift.

Stall

As mentioned in Section 4.1.1, the stall phenomenon is generally associated with a sudden decrease in C_L. It follows that if the stall angle α_s satisfies $\alpha_s + \gamma_\lambda \in I$,the graph of μ_L shall cross the graph of μ_G at an angle in $\varphi \geq \alpha_s + \gamma_\lambda$. This fact is quoted in [68, p.139]: *"In the stall region [...] there may be multiple solutions for C_L. Each of these solutions is possible. The correct solution should be that which maintains the continuity of the angle of attack along the blade span."*

Corrected model

Though $\lim_{\varphi \to 0^+} \mu_G^c(\varphi) = +\infty$, the graph of μ_G^c may no longer be concave when a correction is used for values of a close to 1. Hence possible multiple solutions, e.g. in the case μ_L is affine.

4.3 Solution algorithms

In this section, we focus on the numerical solving of Glauert's model. In the literature, one algorithm is particularly highlighted: it consists in a fixed point iteration applied on the three equations of the model, i.e. either Equations (4.13–4.15) or Equations (4.21–4.23), that we describe in Section 4.3.1. Many articles note that this algorithm is unstable in some cases, preventing convergence and ultimately the solving of the problem. In Section 4.3.2, we propose a new algorithm, for which we prove convergence in a less restrictive framework.

4.3.1 Usual algorithm

The following procedure is broadly used to solve Equations (4.21–4.23). An early presentation of this algorithm is given in [91], but can also be found in several monographs [15, 68, 86, 47]. In particular, the version given in [85] includes a correction for high values of a and corresponds to the following algorithm.

Algorithm 4.1 consists in solving iteratively (4.21), then (4.22) and finally (4.23). Note that the stopping criterium in Step 4 is arbitrary and usually not mentioned in monographs. The convergence of this algorithm is problematic. Instabilities are often observed in practice, as quoted in [85]: *"Note that this set of equations must be solved simultaneously, and in practice, numerical instability can occur"* and *"When local angle of attack is around the stall point, or becomes negative, getting the BEM code to converge can become difficult."*

We also refer to [67] for a specific study on the convergence issues, as well as Appendix 4.A for additional assumptions that guarantee its convergence.

Algorithm 4.1: BEM, usual procedure

Input: Tol > 0, $\alpha \mapsto C_L(\alpha)$, $\alpha \mapsto C_D(\alpha)$, λ, γ_λ, σ_λ, F_λ, $x \mapsto \psi(x)$.
Initial guess: a, a'.
Output: a, a', φ.

Set $err := \text{Tol} + 1$.
Define the functions μ_L^c and μ_D^c by (4.24) using the input data.

while $err > Tol$ **do**

1. Set $\varphi := \operatorname{atan}\left(\frac{1-a}{\lambda(1+a')}\right)$.
2. Set $a = \tau(\varphi)$, i.e. the solution of Equation (4.31).
3. Set $a' = \frac{1-a}{\lambda \sin^2 \varphi}(\mu_L^c \sin\varphi - \mu_D^c \cos\varphi)$.
4. Set $err := \left|\tan\varphi - \frac{1-a}{\lambda(1+a')}\right|$.

end

4.3.2 Alternative algorithms

The corrected model involves many cases that make the description and analysis of a general algorithm difficult. We can however propose some strategies and systematic approaches to compute a solution of different versions of the corrected model.

Fixed-point algorithm associated with (4.33)

If the correction for high values of a is not considered, $\mu_G^c = \mu_G$ and an alternative fixed-point iteration can be proposed: assuming that μ_L^c is differentiable and given an initial value φ^0, we define the sequence $(\varphi^k)_{k\in\mathbb{N}}$ by

$$\varphi^{k+1} = f(\varphi^k), \tag{4.40}$$

with $f(\varphi) = \varphi + \rho\left(\mu_G(\varphi) - \mu_L^c(\varphi) + \tan(\theta_\lambda - \varphi)\mu_D^c(\varphi)\right)$, where $\rho > 0$.

We then have the following result of convergence in the case $\varphi = 0$.

Theorem 4.3.1. *Suppose that* $\max I^+ = \theta_\lambda$, *Assumption 4.1.1 holds and the functions* μ_L^c *and* μ_D^c *are continuously differentiable on* I^+, *with* μ_D^c *increasing. If Equations (4.13–4.15) admit at least one solution in* I^+, *then the sequence* $(\varphi^k)_{k\in\mathbb{N}}$ *defined by Equation* (4.40), *with*

$$\rho = \frac{\varepsilon}{\max\limits_{\varphi\in I^+} {\mu_L^c}' + (1+\tan^2\theta_\lambda)\max\limits_{\varphi\in I^+}\mu_D^c + \sin\theta_\lambda}, \tag{4.41}$$

for some $\varepsilon \in (0,1)$ *and the initial value*

$$\varphi^0 = \theta_\lambda \tag{4.42}$$

converges to $\varphi^\star$*, the largest solution in* $(0, \theta_\lambda]$.

Proof. We first prove that f is increasing. Thanks to the choice of ρ and the concavity of $\varphi \mapsto \mu_G(\varphi)$ on $[0, \theta_\lambda]$, we have:

$$\begin{aligned} f'(\varphi) &= 1 + \rho\left(\mu_G'(\varphi) - {\mu_L^c}'(\varphi) - (1+\tan^2(\theta_\lambda - \varphi))\mu_D^c(\varphi) + \tan(\theta_\lambda - \varphi){\mu_D^c}'(\varphi)\right) \\ &\geq 1 - \rho\left(\max_{\varphi\in I^+} {\mu_L^c}' + (1+\tan^2(\theta_\lambda - \varphi)) \max_{\varphi\in I^+} \mu_D^c + \sin\theta_\lambda\right) \\ &= 1 - \varepsilon \geq 0. \end{aligned}$$

Let us then show that $[\varphi^\star, \theta_\lambda]$ is stable by f. Since f is increasing and $f(\varphi^\star) = \varphi^\star$, it remains to show that $f(\theta_\lambda) \leq \theta_\lambda$. The latter statement follows from:

$$f(\theta_\lambda) = \theta_\lambda + \rho\left(\mu_G(\theta_\lambda) - \mu_L^c(\theta_\lambda)\right) = \theta_\lambda - \rho\mu_L^c(\theta_\lambda) \leq \theta_\lambda.$$

Since $\varphi^0 = \theta_\lambda$, $(\varphi^k)_{k\in\mathbb{N}}$ is bounded and monotonically decreasing, the result follows from the definition of $\varphi^\star$. □

The efficiency of this algorithm depends on the choice of ε and more generally on the value assigned to ρ. In some cases, we can estimate the rate of convergence of $(\varphi^k)_{k\in\mathbb{N}}$.

Theorem 4.3.2. *In addition to the assumptions of Theorem 4.3.1, suppose that*

$$\frac{1}{\lambda} = \tan\theta_\lambda < \min_{\varphi\in[\gamma_\lambda,\theta_\lambda]} {\mu_L^c}'(\varphi) + {\mu_D^c}'(\varphi). \tag{4.43}$$

Then the sequence $(\varphi^k)_{k\in\mathbb{N}}$ *defined by (4.40–4.42) satisfies*

$$|\varphi^k - \varphi^\star| \leq \left(1 - \rho\left(\min_{\varphi\in[\gamma_\lambda,\theta_\lambda]} {\mu_L^c}'(\varphi) + {\mu_D^c}'(\varphi) - \tan\theta_\lambda\right)\right)^k |\theta_\lambda - \rho\mu_L^c(\theta_\lambda)|.$$

Proof. Since we have already proved that $f'(\varphi) \geq 0$ on $[\varphi^\star, \theta_\lambda]$, it remains to determine an upper bound for $f'(\varphi)$. To do this, we use (4.43) and $\mu_G'(\varphi) \leq \mu_G'(0) = \tan\theta_\lambda$ to get:

$$\begin{aligned} f'(\varphi) &= 1 + \rho\left(\mu_G'(\varphi) - {\mu_L^c}'(\varphi) - (1+\tan^2(\theta_\lambda - \varphi))\mu_D^c(\varphi) + \tan(\theta_\lambda - \varphi){\mu_D^c}'(\varphi)\right) \\ &\leq 1 - \rho\left(\min_{\varphi\in[\gamma_\lambda,\theta_\lambda]} {\mu_L^c}'(\varphi) + {\mu_D^c}'(\varphi) - \tan\theta_\lambda\right). \end{aligned}$$

The result follows by induction. □

We can actually obtain a quadratic convergence, i.e. $|\varphi^k - \varphi^\star| \leq \nu|\varphi^0 - \varphi^\star|^{2^k}$ for some $\nu > 0$ by making use of a Newton procedure, that is by replacing ρ given in (4.41) by the sequence $(\rho^k)_{k\in\mathbb{N}}$ defined by

$$\rho^k := -\frac{1}{\mu_G'(\varphi^k) - \mu_L^c{}'(\varphi^k) - (1+\tan^2(\theta_\lambda - \varphi^k))\mu_D^c(\varphi^k) + \tan(\theta_\lambda - \varphi^k)\mu_D^c{}'(\varphi^k)}.$$

Though the term $\mu_G'(\varphi^k)$ can be computed exactly as well as most of the terms in the denominator, the latter expression remains difficult to evaluate since the functions μ_L^c and μ_D^c are only known experimentally, i.e. pointwise.

The general case

If the correction for high values of a is considered and applies, then the framework of Theorem 4.2.8 implies that there exists a solution of the corrected model in I^+. As a consequence, a *bisection algorithm* applied on Equation (4.33) converges to such a solution.

We finally show that the solution found in the case $\psi = 0$ can be used to bracket the solution in the general case of the corrected model.

Lemma 4.3.3. *Keep the assumptions of Corollary 4.2.8, and denote by $\varphi^\star$ a solution of (4.33), when $\psi = 0$. Then (4.33) admits a solution in $(\varphi^\star, \min\{\theta_\lambda, \beta + \gamma_\lambda\}]$ corresponding to a positive lift.*

Proof. Since $\varphi^\star$ satisfies (4.33) with $\psi = 0$, we have:

$$\mu_L^c(\varphi^\star) - \tan(\theta_\lambda - \varphi^\star)\mu_D^c(\varphi^\star) - \mu_G^c(\varphi^\star) = -\frac{\cos\theta_\lambda \sin^2\varphi^\star}{\cos(\theta_\lambda - \varphi^\star)} \frac{\psi\left((\tau(\varphi^\star) - a_c)_+\right)}{(1-\tau(\varphi^\star))^2} \leq 0.$$

The result follows by combining the latter with (4.38) and by applying Corollary 4.2.8. □

4.4 Optimization

The BEM model does not only aim at evaluating the efficiency of a given geometry, but also provides a method to design rotors, that is, to select high-performance parameters γ_λ and c_λ. In this way, monographs generally consider a particular functional, often called *power coefficient* ([68, p.126], [44, p.328] and [86]), which corresponds to the ratio between the energy received and the energy captured.

In this framework, the drag coefficient C_D is taken into account (though partly neglected in the reasoning, as explained hereafter) as well as the Tip loss correction. To the best of our knowledge, no correction related to high values of a, as the one presented in Section 4.1.4, is considered in optimization procedures.

4.4.1 Simplified model and usual design procedure

The design procedure mainly consists in optimizing the power coefficient, defined by the relation

$$C_p(\gamma_\lambda, c_\lambda, \varphi) = \frac{8}{\lambda_{\max}^2} \int_{\lambda_{\min}}^{\lambda_{\max}} \lambda^3 a'(1-a) \left(1 - \frac{C_D}{C_L} \tan^{-1} \varphi\right) d\lambda, \tag{4.44}$$

under the constraints (4.21–4.23), in the sense that only a Tip loss correction is occasionally considered, which actually does not modify the optimization procedure described in [68, p.131-137]. For the sake of completeness, we recall it here keeping the mathematical point of view adopted so far.

The first step consists in defining an angle $\overline{\alpha}$ that minimizes the ratio $\frac{C_D}{C_L}$. Then the coefficient C_D is simply neglected: not only the factor $1 - \frac{C_D}{C_L} \tan^{-1}(\varphi)$ is set to 1 in (4.44), but it also vanishes from the constraints, which now become the simplified model (4.13–4.15). Finally, Lemma 4.2.1 allows us to replace μ_L by μ_G, so that the simplified model (and hence C_p) depends exclusively on φ. Then, the optimization problem is reduced to

$$\max_{\varphi} \; C_p(\varphi) = \frac{8}{\lambda_{\max}} \int_{\lambda_{\min}}^{\lambda_{\max}} \lambda^3 a'(1-a) d\lambda$$
$$\text{s.t.} \begin{cases} \tan\varphi = \dfrac{1-a}{\lambda(1+a')} \\ \dfrac{a}{1-a} = \mu_G(\varphi) \dfrac{\cos\varphi}{\sin^2\varphi} \\ \dfrac{a'}{1-a} = \mu_G(\varphi) \dfrac{1}{\lambda \sin\varphi}. \end{cases}$$

Recalling that $\mu_G(\varphi) = \sin(\varphi)\tan(\theta_\lambda - \varphi)$, we can express a and a' as functions of φ, namely

$$1 - a = \frac{\sin\varphi \cos(\theta_\lambda - \varphi)}{\sin\theta_\lambda}, \quad a' = \frac{\sin\varphi \sin(\theta_\lambda - \varphi)}{\cos\theta_\lambda},$$

so that C_p reads

$$C_p = \frac{8}{\lambda_{\max}^2} \int_{\lambda_{\min}}^{\lambda_{\max}} \frac{\lambda^2(1+\lambda^2)}{2} \sin^2\varphi \sin\left(2(\theta_\lambda - \varphi)\right) d\lambda.$$

It then remains to optimize $\varphi \mapsto \sin^2\varphi \sin(2(\theta_\lambda - \varphi))$, for $\varphi \in [0, \theta_\lambda]$. An easy computation shows that the maximum is attained at $\varphi^* = \frac{2}{3}\theta_\lambda$. Finally, the design parameters $\gamma_\lambda^* = \gamma_\lambda(\varphi^*)$ and $c_\lambda^* = c_\lambda(\varphi^*)$ can then be computed from Equations (4.5) and (4.27), to get

$$\gamma_\lambda^* = \varphi^* - \overline{\alpha}, \quad c_\lambda^* = \frac{8\pi r \mu_G(\varphi^*)}{B C_L(\overline{\alpha})}. \tag{4.45}$$

4.4.2 Asymptotical analysis of the corrected model

The optimization problem associated with the corrected model, in the case of Glauert empirical correction ($F_\lambda(\varphi) = 1$), is given by

$$\begin{aligned} \max_{\varphi} \quad & J(\varphi) \;= a'(1-a)\left(1 - \frac{\mu_D(\varphi)}{\mu_L(\varphi)}\tan^{-1}\varphi\right) \\ \text{s.t.} \quad & \begin{cases} \mu_L(\varphi) = \mu_G^c(\varphi) + \tan(\theta_\lambda - \varphi)\mu_D(\varphi) \\ a = \tau(\varphi) \\ a' = \dfrac{1-\tau(\varphi)}{\lambda \sin^2\varphi}\left(\mu_L(\varphi)\sin\varphi - \mu_D(\varphi)\cos\varphi\right). \end{cases} \end{aligned} \tag{4.46}$$

It is clear that its solving becomes more complicated than in the simplified case. However, what if we can identify regimes where the correction is not active at the optimum? We provide a asymptotic result in a neighborhood of $\varphi = 0$ where we discuss this situation.

Theorem 4.4.1. *Suppose that Lemma 4.2.3 holds and the functions μ_D and $\psi\left((a-a_c)_+\right)$ are differentiable. Then, in a neighborhood $[0,\delta)$ of $\varphi = 0$, the cost functional satisfies*

$$J(\varphi) = \frac{\lambda}{\mu_D(0)}\varphi^2 + o_{\varphi=0}(\varphi^2).$$

Moreover, if g is decreasing and $\varphi_c \in [0,\delta)$, then the optimal solution belongs to the region where the correction $\psi((a-a_c)_+)$ is not active.

Before stating the proof, an improved version of Lemma 4.2.6 is required.

Lemma 4.4.2. *Under the assumptions of Lemma 4.2.6, we have*

$$(1-\tau(\varphi))^2 = \frac{\psi(1-a_c)}{\mu_D(0)}\varphi^3 + \frac{\lambda\sqrt{\psi(1-a_c)}}{\mu_D^{3/2}(0)}\varphi^{3+1/2} + o_{\varphi=0}(\varphi^{3+1/2}).$$

Proof. Consider the following expansions that hold in a neighborhood of $\varphi = 0$:

$$\begin{aligned} \frac{\sin\theta_\lambda \sin\varphi}{\cos(\theta_\lambda - \varphi)}\psi\left((a-a_c)_+\right) &= \frac{\psi(1-a_c)}{\lambda}\varphi - \frac{\psi(1-a_c)}{\lambda^2}\varphi^2 + o_{\varphi=0}(\varphi^2), \\ g(\varphi) &= \frac{\mu_D(0)}{\lambda}\frac{1}{\varphi^2} + \Big(\frac{1}{\lambda} + \frac{\mu_D'(0)}{\lambda} - \frac{\mu_D(0)}{\lambda^2}\Big)\frac{1}{\varphi} + o_{\varphi=0}(\frac{1}{\varphi}), \end{aligned}$$

where $\frac{1}{\lambda} = \tan\theta_\lambda$. Note that the first can be explained as follows: in a neighborhood of $a = 1$ (or equivalently around $\varphi = 0$), due to Lemma 4.2.6 we have

$$\begin{aligned} \psi\left((a-a_c)_+\right) &= \psi(1-a_c) + \left.\frac{d\psi((a-a_c)_+)}{da}\right|_{a=1}(a-1) + o_{a=1}(a-1) \\ &= \psi(1-a_c) + \left(\left.\frac{d\psi((a-a_c)_+)}{da}\right|_{a=1}\sqrt{\frac{\psi(1-a_c)}{\mu_D(0)}}\right)\varphi^{3+1/2} + o_{\varphi=0}(\varphi^{3+1/2}). \end{aligned}$$

Defining $\nu(\varphi) = 1 - \tau(\varphi)$, we multiply Equation (4.31) by $\nu^2(\varphi)$ and then use the previous expansions. After rearranging terms, we have

$$\frac{\mu_D(0)}{\lambda}\frac{\nu^2(\varphi)}{\varphi^2} = \nu(\varphi) - \nu^2(\varphi) + \frac{\psi(1-a_c)}{\lambda}\varphi - \frac{\psi(1-a_c)}{\lambda^2}\varphi^2 + o_{\varphi=0}(\varphi^2) \\ - \left[\left(\frac{1}{\lambda} + \frac{\mu_D'(0)}{\lambda} - \frac{\mu_D(0)}{\lambda^2}\right) + o_{\varphi=0}(1)\right]\frac{\nu^2(\varphi)}{\varphi}.$$

We define a function $\xi(\varphi)$ such that $\nu^2(\varphi) = \ell\varphi^3(1+\xi(\varphi))$, with $\ell = \frac{\psi(1-a_c)}{\mu_D(0)}$. Plugging this expression on the left-hand side, and using on the right-hand side the expansions $\nu(\varphi) = \sqrt{\ell}\varphi^{3/2} + o_{\varphi=0}(\varphi^{3/2})$ and $\nu^2(\varphi) = \ell\varphi^3 + o_{\varphi=0}(\varphi^3)$, lead to

$$\frac{\psi(1-a_c)}{\lambda}\varphi\xi(\varphi) = \sqrt{\ell}\varphi^{3/2} + o_{\varphi=0}(\varphi^{3/2}) - \frac{\psi(1-a_c)}{\lambda^2}\varphi^2 \\ - \left[\ell\left(\frac{1}{\lambda} + \frac{\mu_D'(0)}{\lambda} - \frac{\mu_D(0)}{\lambda^2}\right) + o_{\varphi=0}(1)\right]\varphi^2.$$

Since $\varphi^2 = o_{\varphi=0}(\varphi^{3/2})$, we can merge all the quadratic terms with $o_{\varphi=0}(\varphi^{3/2})$. Dividing by φ, we finally get

$$\xi(\varphi) = \frac{\lambda\sqrt{\ell}}{\psi(1-a_c)}\varphi^{1/2} + o_{\varphi=0}(\varphi^{1/2}). \qquad \square$$

Proof of Theorem 4.4.1. We begin by deriving an explicit expression for the cost functional. After replacing the constraints associated with a and a' into $J(\varphi)$, we obtain

$$J(\varphi) = \frac{\left[(1-\tau(\varphi))^2(\mu_L(\varphi)\sin\varphi - \mu_D(\varphi)\cos\varphi)\right]^2}{\lambda(1-\tau(\varphi))^2\mu_L(\varphi)\sin^3\varphi}. \tag{4.47}$$

Note that Equation (4.34) allows us to write explicitely the numerator. Indeed, we have

$$\begin{aligned}(1-\tau(\varphi))^2&(\mu_L(\varphi)\sin\varphi - \mu_D(\varphi)\cos\varphi) \\ &= (1-\tau(\varphi))^2\sin^2\varphi\tan(\theta_\lambda - \varphi) \\ &\quad + \frac{\cos\theta_\lambda}{\cos(\theta_\lambda-\varphi)}\left(\sin^3\varphi\,\psi((\tau(\varphi)-a_c)_+) - (1-\tau(\varphi))^2\mu_D(\varphi)\right).\end{aligned} \tag{4.48}$$

In order to determine a regime where the optimum of $J(\varphi)$ is attained in $[0, a_c)$, i.e. when the correction does not apply, we study the asympotical behavior of $J(\varphi)$ around $\varphi = 0$ (or equivalently, $a = 1$). With the help of Lemma 4.2.6, we can expand the first terms of the right-hand side of (4.48), to get

$$(1-\tau(\varphi))^2\sin^2\varphi\tan(\theta_\lambda-\varphi) = \frac{\psi(1-a_c)}{\lambda\mu_D(0)}\varphi^5 + o_{\varphi=0}(\varphi^5),$$
$$\frac{\cos\theta_\lambda}{\cos(\theta_\lambda-\varphi)} = 1 + o_{\varphi=0}(1),$$
$$\sin^3\varphi\,\psi((a-a_c)_+) = \psi(1-a_c)\cdot\varphi^3 + o_{\varphi=0}(\varphi^4).$$

However, applying Lemma 4.2.6 to the last term in (4.48) leads to a right-hand side of order $o_{\varphi=0}(\varphi^3)$, and then $J(\varphi) = o_{\varphi=0}(\varphi)$. Instead, we use Lemma 4.4.2, which brings

$$\begin{aligned}(1-\tau(\varphi))^2\mu_D(\varphi) &= \Big(\frac{\psi(1-a_c)}{\mu_D(0)}\varphi^3 + \frac{\lambda\sqrt{\psi(1-a_c)}}{\mu_D^{3/2}(0)}\varphi^{3+1/2} + o_{\varphi=0}(\varphi^{3+1/2})\Big)\\ &\qquad \times (\mu_D(0) + \mu_D'(0)\cdot\varphi + o_{\varphi=0}(\varphi))\\ &= \psi(1-a_c)\cdot\varphi^3 + \frac{\lambda\sqrt{\psi(1-a_c)}}{\sqrt{\mu_D(0)}}\varphi^{3+1/2} + o_{\varphi=0}(\varphi^{3+1/2})\end{aligned}$$

and then, replacing all these expressions into (4.48) and using that $\varphi^4 = o(\varphi^{3+1/2})$, yields

$$(1-\tau(\varphi))^2(\mu_L(\varphi)\sin\varphi - \mu_D(\varphi)\cos\varphi) = \frac{\lambda\sqrt{\psi(1-a_c)}}{\sqrt{\mu_D(0)}}\varphi^{3+1/2} + o_{\varphi=0}(\varphi^{3+1/2}). \quad (4.49)$$

From the previous calculations we also obtain

$$(1-\tau(\varphi))^2\mu_L(\varphi)\sin\varphi = \psi(1-a_c)\cdot\varphi^3 + o_{\varphi=0}(\varphi^3). \quad (4.50)$$

Now we can expand the cost functional by replacing (4.49) and (4.50) in (4.47), to get

$$J(\varphi) = \frac{\left[\frac{\lambda\sqrt{\psi(1-a_c)}}{\sqrt{\mu_D(0)}}\varphi^{3+1/2} + o_{\varphi=0}(\varphi^{3+1/2})\right]^2}{\lambda\left(\psi(1-a_c)\cdot\varphi^3 + o_{\varphi=0}(\varphi^3)\right)(\varphi + o_{\varphi=0}(\varphi^2))^2} = \frac{\lambda}{\mu_D(0)}\varphi^2 + o_{\varphi=0}(\varphi^2).$$

Finally, we denote by $[0,\delta)$ the neighborhood where all these computations hold. Since the cost functional is increasing in this interval, necessarily the optimum φ^* is attained outside. Then, if exists $\varphi_c \in [0,\delta)$ such that $\tau(\varphi_c) = a_c$, we have $\tau(\varphi^*) \leq \tau(\delta) \leq a_c$, due to the fact that τ is a decreasing function. It follows that a correction for high values of a is not required. □

4.5 Numerical experiments

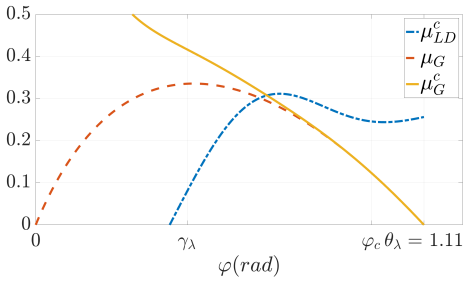
In this section, we investigate on a practical case the performance of the algorithms presented in Section 4.3 and study numerically the design optimization problem considered in Section 4.4.

4.5.1 A practical example

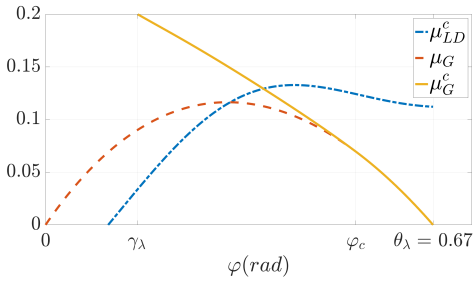
We consider a turbine consisting of three blades, designed with a NACA 4415 profile and three different blade elements of rotation speeds $\lambda_1 = 0.5$, $\lambda_2 = 1.25$, and $\lambda_3 = 3$, respectively. In all these cases, we set $\gamma_\lambda = \gamma_\lambda^*$ and $c_\lambda = c_\lambda^*$, i.e. we use the optimal values of the simplified model given by (4.45) (where $\bar{\alpha} = 0.105$ radians):

λ	$\gamma^\star$	$c_\lambda^\star$
$\lambda_1 = 0.5$	0.432028	0.308838
$\lambda_2 = 1.25$	0.159495	0.320211
$\lambda_3 = 3$	-0.0392376	0.18377

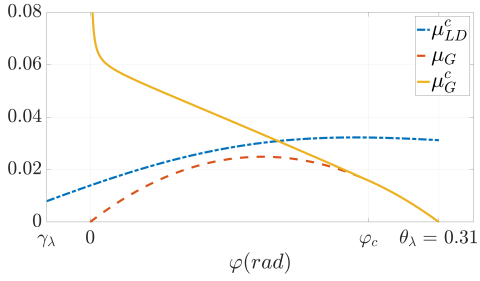
We use the correction from Wilson *et al* and Spera, see Table 4.1. In this example, $I^+ = (0, \theta_\lambda]$ for all elements. Graphs of the functions $\mu_{LD}^c : \varphi \mapsto \mu_L^c(\varphi) - \tan(\theta_\lambda - \varphi)\mu_D^c(\varphi)$, μ_G^c and μ_G are presented below.



(a) $\lambda_1 = 0.5$



(b) $\lambda_2 = 1.25$



(c) $\lambda_3 = 3$

Figure 4.2: Graphs of of the functions μ_{LD}^c, μ_G^c and μ_G for different values of λ. Note that these figures are similar to the scheme given in [68, Figure 3.27, p.126].

4.5.2 Solution algorithms

We first compare the various solution algorithms, in terms of iterations. After having computed accurately a numerical solution ϕ^∞, we measure the number of iterations required to reach an error satisfying

$$err^k := |\phi^k - \phi^\infty| \leq \text{Tol} = 10^{-5}.$$

We consider the three algorithms presented in Section 4.3, that is, the usual aglorithm Algorithm 4.1, the bisection algorithm detailed in Section 4.3.2 and the new fixed-point algorithm corresponding to the iteration (4.40). For the latter, we use for ρ the value given by Equation (4.41) (though we include in our test a correction on α), with $\varepsilon = 1$.

λ	Algorithm 4.1	Bisection	Fixed-point Algorithm
$\lambda_1 = 0.5$	8	17	19
$\lambda_2 = 1.25$	12	16	8
$\lambda_3 = 3$	15	15	8

Table 4.2: Number of iterations required to solve Equations (4.21–4.23)

We see in the table above that though we do not use an optimal value for ρ, the convergence of our new fixed-point algorithm converges faster than the others. To evaluate the effect of a change in ρ, we plot the values of err^k with respect to the iterations in the case where $\lambda = \lambda_2$. The results are presented in Figure 4.3.

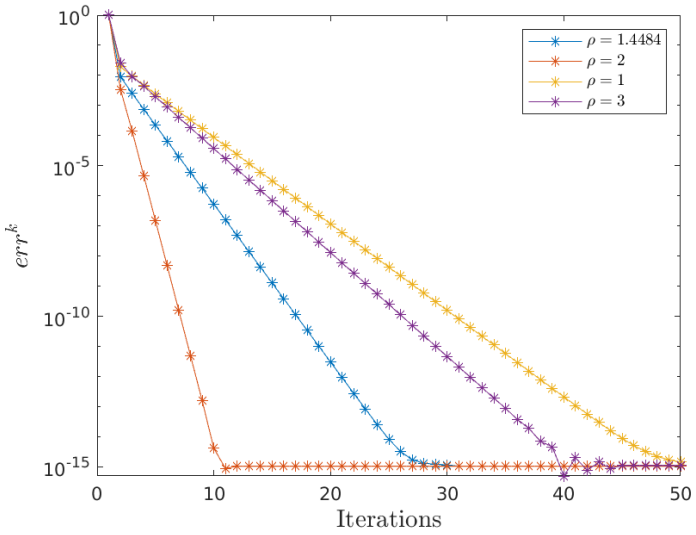


Figure 4.3: Convergence of the new fixed-point algorithm for various values of ρ

The value $\rho = 1.4484$ is the one given by Equation (4.41) for $\varepsilon = 1$. The test confirms that this value is not optimal, and can be easily improved. However, using greater or smaller values can deteriorate the convergence, as is the case for $\rho = 1$ or $\rho = 3$.

4.5.3 Optimization

In this test, we work with three values $\widetilde{\lambda_1} = 1.05$, $\widetilde{\lambda_2} = 2.96$ and $\widetilde{\lambda_3} = 4.88$, corresponding to three elements of the actual turbine used by *Hydrotube Energy*.

We use a gradient method to compute a solution of (4.46) and compare it to the (explicit) optimal solution associated with the simplified model given by Equations (4.45). For the sake of completeness, we detail now the optimization method we use. In particular, we recall here how the introduction of an adjoint variable enables to compute the gradient of J. In this last section, we denote by C_L' and C_D' the derivatives of C_L and C_D respectively, and omit the dependence of μ_L and μ_D and their derivatives on the variable φ for the sake of simplicity.

We define the Lagrangian of Problem 4.46 by

$$\begin{aligned}
&\mathcal{L}(\varphi, a, a', p_1, p_2, p_3, c_\lambda, \gamma_\lambda) \\
&\qquad := J(\varphi, a, a', c_\lambda, \gamma_\lambda) - p_1 \cdot (\mu_L(\varphi) - \tan(\theta_\lambda - \varphi)\mu_D(\varphi) - \mu_G^c(\varphi)) \\
&\qquad\quad - p_2 \cdot \left(\frac{a}{1-a} - \frac{1}{\sin^2\varphi}(\mu_L^c(\varphi)\cos\varphi + \mu_D^c(\varphi)\sin\varphi) + \frac{\psi\left((a-a_c)_+\right)}{(1-a)^2} \right) \\
&\qquad\quad - p_3 \cdot \left(\frac{a'}{1-a} - \frac{1}{\lambda\sin^2\varphi}(\mu_L^c(\varphi)\sin\varphi - \mu_D^c(\varphi)\cos\varphi \right),
\end{aligned}$$

where p_1, p_2 and p_3 are the Lagrange multipliers associated with the constraints (4.21–4.23).

The optimality system is obtained by cancelling all the partial derivatives of $\mathcal{L}$. Differentiating $\mathcal{L}$ with respect to p_1, p_2 and p_3 and equating the resulting terms to zero gives the corrected model Equations (4.21–4.23), that can be solved using the algorithms presented in Section 4.3. Setting the derivatives of $\mathcal{L}$ with respect to (φ, a, a') to zero gives rise to the linear system

$$M \cdot p = b \tag{4.51}$$

where $p := (p_1 \;\; p_2 \;\; p_3)^\top$ is the Lagrange multiplier vector, and

$$M := \begin{bmatrix}
\frac{1}{\cos^2\varphi} & \frac{\mu_D - \frac{\partial\mu_L}{\partial\varphi}}{\sin\varphi\tan\varphi} + \frac{\mu_L(1-2\tan^{-2}\varphi) - \frac{\partial\mu_D}{\partial\varphi}}{\sin\varphi} & \frac{\mu_L + \frac{\partial\mu_D}{\partial v}}{\lambda\sin\varphi\tan\varphi} - \frac{\frac{\partial\mu_L}{\partial\varphi} + \mu_D(1+2\tan^{-2}\varphi)}{\lambda\sin\varphi} \\
-\frac{1}{\lambda(1+a')} & \frac{1+\psi'((a-a_c)_+)}{(1-a)^2} + \frac{2\psi((a-a_c)_+)}{(1-a)^3} & \frac{a'}{(1-a)^2} \\
\frac{1-a}{\lambda(1+a')^2} & 0 & \frac{1}{1-a}
\end{bmatrix}$$

$$b := \begin{pmatrix} a'(1-a)\frac{C_L'(\varphi-\gamma_\lambda)C_D(\varphi-\gamma_\lambda)-C_D'(\varphi-\gamma_\lambda)C_L(\varphi-\gamma_\lambda)}{C_L(\varphi-\gamma_\lambda)^2\tan\varphi} + \frac{C_D(\varphi-\gamma_\lambda)}{C_L(\varphi-\gamma_\lambda)\sin^2\varphi} \\ -a'\frac{1-C_D(\varphi-\gamma_\lambda)}{C_L(\varphi-\gamma_\lambda)\tan\varphi} \\ (1-a)\frac{1-C_D(\varphi-\gamma_\lambda)}{C_L(\varphi-\gamma)\tan\varphi} \end{pmatrix}.$$

We are in a position to detail how the gradient can be computed. Fix the values of the pair $(\gamma_\lambda, c_\lambda)$. If φ, a, a' are the corresponding solutions of Equations (4.21–4.23) and p is the associated solution of Equation (4.51), then the gradient $\nabla J(\gamma_\lambda, c_\lambda)$ of $J(\gamma_\lambda, c_\lambda)$ is given by

$$\nabla J(\gamma_\lambda, c_\lambda) = \begin{pmatrix} \frac{\partial \mathcal{L}}{\partial \gamma_\lambda} & \frac{\partial \mathcal{L}}{\partial c_\lambda} \end{pmatrix}^\top, \tag{4.52}$$

where

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \gamma_\lambda} &= a'(1-a)\frac{C_D'(\varphi-\gamma_\lambda)C_L(\varphi-\gamma_\lambda) - C_L'(\varphi-\gamma_\lambda)C_D(\varphi-\gamma_\lambda)}{C_L^2(\varphi-\gamma_\lambda)\tan\varphi} \\ &\quad - p_2 \cdot \frac{1}{\sin^2\varphi}\left(\frac{\partial \mu_L}{\partial \varphi}\cos\varphi + \frac{\partial \mu_D}{\partial \varphi}\sin\varphi\right) - p_3 \cdot \frac{1}{\lambda\sin^2\varphi}\left(\frac{\partial \mu_L}{\partial \varphi}\sin\varphi - \frac{\partial \mu_D}{\partial \varphi}\cos\varphi\right), \\ \frac{\partial \mathcal{L}}{\partial c_\lambda} &= p_2 \cdot \frac{1}{\sin^2\varphi}\left(\frac{\partial \mu_L}{\partial c_\lambda}\cos\varphi + \frac{\partial \mu_D}{\partial c_\lambda}\sin\varphi\right) + p_3 \cdot \frac{1}{\lambda\sin^2\varphi}\left(\frac{\partial \mu_L}{\partial c_\lambda}\sin\varphi - \frac{\partial \mu_D}{\partial c_\lambda}\cos\varphi\right). \end{aligned}$$

The optimization algorithm is then:

Algorithm 4.2: Numerical optimization

Input: Tol > 0, $\kappa > 0$, $\alpha \mapsto C_L(\alpha)$, $\alpha \mapsto C_D(\alpha)$, λ, $x \mapsto \psi(x)$.
Initial guess: $\gamma_\lambda, c_\lambda$.
Output: $\gamma_\lambda, c_\lambda$.

Set $err := \text{Tol} + 1$.
Define the functions μ_L^c and μ_D^c by (4.24) using the input data.

while $err > Tol$ **do**

1. Set φ, a, a' as the solutions of Equations (4.21–4.23).
2. Set p as the solution of Equation (4.51).
3. Compute the gradient $\nabla J(\gamma_\lambda, c_\lambda)$ given by Equation (4.52).
4. Update $\begin{pmatrix} c_\lambda \\ \gamma_\lambda \end{pmatrix} = \begin{pmatrix} c_\lambda \\ \gamma_\lambda \end{pmatrix} + \kappa \nabla J(\gamma_\lambda, c_\lambda)$,
5. Set $err := \|\nabla J(\gamma_\lambda, c_\lambda)\|$.

end

We apply this algorithm using as initial guess the solution (4.45) associated with the simplified model. The results are presented in Table 4.3. We see that the gradient procedure enables to improve significantly the values of J, namely by 7.4918%, 14.267% and 19.907% for the three elements under consideration. Notice that in the three cases, the correction is activated, i.e. $a > a_c$, though the value of a is slightly lower than with the parameters values given by (4.45).

	J		c_λ		γ_λ	
	Simplified model	Corrected model	Simplified model	Corrected model	Simplified model	Corrected model
$\widetilde{\lambda_1} = 1.05$	0.117795	0.126620	0.370531	0.344068	0.409762	0.217279
$\widetilde{\lambda_2} = 2.96$	0.015722	0.017965	0.196778	0.193098	0.116262	-0.034469
$\widetilde{\lambda_3} = 4.88$	0.005365	0.006433	0.125542	0.121822	0.032674	-0.102584

Table 4.3: Optimal values obtained with Algorithm 4.2

4.6 Perspectives

This work focuses mainly on providing conditions that ensure the existence of solutions for the BEM model. Several questions regarding the convergence of solving algorithms, as well as the associated optimization problem, remain to be answered.

Concerning the former, extending the proof to a more general framework is desirable, in particular for the simplified case, since its analysis rely on a fixed point argument. The optimization problem also offers several possibilities. An asymptotic analysis gives a general idea of the optimal solution behavior, however the required assumptions seem very restrictive. It could be useful to prove a similar result by using optimization tools, to confirm when the correction for high values of a is needed. If it is not the case, a next step concerns the negligibility of C_D, focusing on the difference between the optimal solution and the obtained in the simplified case. Finally, the question of multiple optima in the corrected model remains open. A systematic use of numerical experiments on various cases could provide first answers.

Appendix 4.A Convergence in the simplified case

In the simplified case, Algorithm 4.1 can be summarized as the calculation of the terms of a sequence $(\varphi^k)_{k\in\mathbb{N}}$ associated with the following recursion:

$$\varphi^{k+1} := \widetilde{f}(\varphi^k), \tag{4.53}$$

with $\widetilde{f}(x) := \frac{\pi}{2} - \operatorname{atan}\left(\lambda + \mu_L(x)h(x)\right)$ and $h(x) := \frac{\lambda \tan^{-1} x + 1}{\sin x}$.

The stability can actually be obtained in the simplified case with additional assumptions.

Lemma 4.A.1. *Suppose that μ_L is defined, positive and increasing on $[\gamma_\lambda, \theta_\lambda]$, and that*

$$\mu_L(\theta_\lambda) \leq \mu_G(\gamma_\lambda). \tag{4.54}$$

If the initial value φ^0 belongs to $[\gamma_\lambda, \theta_\lambda]$, then the sequence defined by (4.53) satisfies:

$$\forall k \in \mathbb{N},\ \varphi^k \in [\gamma_\lambda, \theta_\lambda].$$

Proof. The result can be obtained by induction. Assume that for some $k \in \mathbb{N}$, $\varphi^k \in [\gamma_\lambda, \theta_\lambda]$. Because of (4.53), we have

$$\tan^{-1} \varphi^{k+1} := \lambda + \mu_L(\varphi^k)h(\varphi^k),$$

so that, since $\mu_L \geq 0$ and by Definition (4.25) of θ_λ, then $\varphi^{k+1} \leq \theta_\lambda$. On the other hand, thanks to the assumption (4.54), we obtain

$$\tan^{-1} \varphi^{k+1} \leq \lambda + \mu_L(\theta_\lambda)h(\gamma_\lambda).$$

Because of (4.54), the left-hand side of the last inequality is bounded by $\tan^{-1} \gamma_\lambda$. The result follows. □

To get a sufficient condition for convergence, the last result must be completed by a contraction property. This is the object of the following result.

Lemma 4.A.2. *Suppose that μ_L is differentiable and denote by μ'_L its derivative. The derivative of $\widetilde{f}$ satisfies*

$$-\frac{\max\limits_{\varphi\in I} \mu'_L h(\gamma_\lambda)}{1+\lambda^2} \leq \widetilde{f}'(\varphi) \leq \frac{\max\limits_{\varphi\in I} \mu_L |h'(\gamma_\lambda)|}{1+\lambda^2}.$$

Proof. Differentiating $\widetilde{f}$, we find that

$$\widetilde{f}'(\varphi) = \frac{-1}{1 + \left(\lambda + \mu_L(\varphi)h(\varphi)\right)^2}\left(\mu'_L(\varphi)h(\varphi) + \mu_L(\varphi)h'(\varphi)\right).$$

For $\varphi \in (\gamma_\lambda, \theta_\lambda)$, $\mu_L'(\varphi)h(\varphi) \geq 0$ and $\mu_L(\varphi)h'(\varphi) \leq 0$, so that

$$\frac{-1}{1+(\lambda+\mu_L(\varphi)h(\varphi))^2}\mu_L'(\varphi)h(\varphi) \leq f'(\varphi) \leq \frac{-1}{1+(\lambda+\mu_L(\varphi)h(\varphi))^2}\mu_L(\varphi)h'(\varphi).$$

The result then follows from the fact that $\mu_L(\varphi)h(\varphi) \geq 0$, h and h' is decreasing on $(\gamma_\lambda, \theta_\lambda)$. □

We are now in the position to obtain a conditional convergence result, the condition being stated in the next assumption.

Theorem 4.A.3. *In addition to the assumptions of Lemma 4.A.1, suppose that μ_L is differentiable and satisfies*

$$\frac{\max_I \mu_L' h(\gamma_\lambda)}{1+\lambda^2} \leq 1 \tag{4.55}$$

$$\frac{\max_I \mu_L |h'(\gamma_\lambda)|}{1+\lambda^2} \leq 1. \tag{4.56}$$

Then, if φ^0 belongs to $[\gamma_\lambda, \theta_\lambda]$, the sequence $(\varphi^k)_{k\in\mathbb{N}}$ defined by Equation (4.53) converges to the unique solution of Equation (4.27).

Proof. As a consequence of Lemma 4.A.1, the function $\widetilde{f}$ maps $[\gamma_\lambda, \theta_\lambda]$ onto itself. From Equations (4.55), (4.56) and Lemma 4.A.2 we deduce $\widetilde{f}$ is contracting. The result follows from the Banach Fixed Point Theorem. □

References

[1] J. Ackermann. Der entwurf linearer regelungssysteme im zustandsraum. *Regelungstechnik*, 20:297–300, 1972.

[2] J. Ackermann. On the synthesis of linear control systems with specified characteristics. *Automatica*, 13:89–94, 1977.

[3] C. Afri, V. Andrieu, L. Bako, and P. Dufour. State and parameter estimation: A nonlinear Luenberger observer approach. *IEEE Transactions on Automatic Control*, 62(2):973–980, 2017.

[4] G. Allaire. *Numerical analysis and optimization: An introduction to mathematical modelling and numerical simulation*. Numerical Mathematics and Scientific Computation. Oxford University Press, oup edition, 2007.

[5] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, New York, 2000.

[6] D. Auroux and J. Blum. Back and forth nudging algorithm for data assimilation problems. *Comptes rendus de l'Académie des sciences. Série I, Mathématique*, 340:873–878, 2005.

[7] G. Bal. Parallelization in time of (stochastic) ordinary differential equations. Preprint, 2003.

[8] H. T. Banks and K. Kunisch. *Estimation Techniques for Distributed Parameter Systems*. Systems & Control: Foundations & Applications. Birkhäuser Basel, 1989.

[9] S. Bartels. Total variation minimization with finite elements: convergence and iterative solution. *SIAM Journal on Numerical Analysis*, 50(3):1162–1180, 2012.

[10] E. Barthélemy. Nonlinear shallow water theories for coastal waves. *Surveys in Geophysics*, 25(3):315–337, 2004.

[11] H. Barucq, T. Chaumont-Frelet, and C. Gout. Stability analysis of heterogeneous Helmholtz problems and finite element solution based on propagation media approximation. *Mathematics of Computation*, 86(307):2129–2157, 2017.

[12] R. Bass and I. Gura. High-order system design via state-space considerations. In *Joint Automatic Control Conference*, volume 3, pages 311–319, New York, 1965.

[13] A. Bastide, P.-H. Cocquet, and D. Ramalingom. Penalization model for Navier-Stokes-Darcy equation with application to porosity-oriented topology optimization. *Mathematical Models and Methods in Applied Sciences (M3AS)*, 28(8):1481–1512, 2018.

[14] A. Bouharguane and B. Mohammadi. Minimization principles for the evolution of a soft sea bed interacting with a shallow. *International Journal of Computational Fluid Dynamics*, 26(3):163–172, 2012.

[15] E. Branlard. *Wind turbine aerodynamics and vorticity-based methods: fundamentals and recent applications*. Research topics in wind energy. Springer, Cham, 2017.

[16] O. Bristeau and J. Sainte-Marie. Derivation of a non-hydrostatic shallow water model; comparison with Saint-Venant and Boussinesq systems. *Discrete and Continuous Dynamical Systems - Series B (DCDS-B)*, 10(4):733–759, 2008.

[17] D. Brown, D. Gallistl, and D. Peterseim. Multiscale Petrov-Galerkin method for high-frequency heterogeneous Helmholtz equations. In M. Griebel and M. Schweitzer, editors, *Meshfree Methods for Partial Differential Equations VIII*, Springer Lecture notes in computational science and engineering 115, pages 85–115. Springer, 2017.

[18] P. Bělík and M. Luskin. Approximation by piecewise constant functions in a BV metric. *Mathematical Models and Methods in Applied Sciences*, 13(3):373–393, 2003.

[19] M. L. Buhl, Jr. New empirical relationship between thrust coefficient and induction factor for the turbulent windmill state. *Technical Report NREL/TP-500-36834*, 8 2005.

[20] T. Burton, D. Sharpe, N. Jenkins, and E. Bossanyi. *The Wind Energy Handbook*, volume 1. John Wiley and Sons, Ltd, 01 2001.

[21] Z. Chen and J. Zou. An augmented Lagrangian method for identifying discontinuous parameters in elliptic systems. *SIAM Journal on Control and Optimization*, 37(3):892–910, 1999.

[22] T. Coleman and Y. Li. On the convergence of interior-reflective Newton methods for nonlinear minimization subject to bounds. *Mathematical Programming*, 67(1):189–224, 1994.

[23] T. Coleman and Y. Li. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM Journal of Optimization*, 6(2):418–445, 1996.

[24] P. Courtier, J.-N. Thépaut, and A. Hollingsworth. A strategy for operational implementation of 4d-var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, 120(519):1367–1387, 1994.

[25] J. Dalphin and R. Barros. Shape optimization of a moving bottom underwater generating solitary waves ruled by a forced KdV equation. *Journal of Optimization Theory and Applications*, 180(2):574–607, 2019.

[26] L. D'Amore and R. Cacciapuoti. DD-DA PinT-based model: A domain decomposition approach in space and time, based on parareal, for solving the 4D-Var data assimilation model. *ArXiv e-prints*, 2018.

[27] A. Decoene, L. Bonaventura, E. Miglio, and F. Saleri. Asymptotic derivation of the section-averaged shallow water equations for river hydraulics. *Mathematical Models and Methods in Applied Sciences (M3AS)*, 19:387–417, 2009.

[28] D. Dutykh and H. Kalisch. Boussinesq modeling of surface waves due to underwater landslides. *Nonlinear Processes in Geophysics*, 20:267–285, 2013.

[29] S. Engblom. Parallel in time simulation of multiscale stochastic chemical kinetics. *Multiscale Modeling & Simulation*, 8(1):46–68, 2009.

[30] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag New York, 2004.

[31] S. Esterhazy and J. M. Melenk. On stability of discretizations of the Helmholtz equation. In *Numerical analysis of multiscale problems*, volume 83 of *Lecture Notes in Computational Science and Engineering*, pages 285–324. Springer Verlag, Berlin, Heidelberg, 2012.

[32] W. Froude. On the elementary relation between pitch, slip and propulsive efficiency. *Trans. Roy. Inst. Naval Arch.*, 19(47):47–57, 1878.

[33] M. J. Gander. Schwarz methods over the course of time. *ETNA. Electronic Transactions on Numerical Analysis [electronic only]*, 31:228–255, 2008.

[34] M. J. Gander. 50 years of time parallel time integration. In T. Carraro, M. Geiger, S. Körkel, and R. Rannacher, editors, *Multiple Shooting and Time Domain Decomposition Methods*, volume 9 of *Contributions in Mathematical and Computational Sciences*, pages 69–113. Springer, 2015.

[35] M. J. Gander and S. Güttel. ParaExp: A parallel integrator for linear initial-value problems. *SIAM Journal on Scientific Computing*, 35(3):C123–C142, 2013.

[36] M. J. Gander and E. Hairer. Nonlinear convergence analysis for the parareal algorithm. In O. B. Widlund and D. E. Keyes, editors, *Domain Decomposition Methods in Science and Engineering XVII*, volume 60 of *Lecture Notes in Computational Science and Engineering*, pages 45–56. Springer, 2008.

[37] M. J. Gander and L. Halpern. Méthodes de décomposition de domaines notions de base. *Techniques de l'ingénieur Analyse numérique des équations différentielles et aux dérivées partielles*, base documentaire : 42620210.(ref. article : af1375), 2012.

[38] M. J. Gander, F. Kwok, and H. Zhang. Multigrid interpretations of the parareal algorithm leading to an overlapping variant and MGRIT. *Computing and Visualization in Science*, 19(3):59–74, 2018.

[39] M. J. Gander and S. Vandewalle. Analysis of the parareal time-parallel time-integration method. *SIAM Journal on Scientific Computing*, 29:556–578, 2007.

[40] J.-F. Gerbeau and B. Perthame. Derivation of viscous Saint-Venant system for laminar shallow water; numerical validation. *Discrete and Continuous Dynamical Systems - Series B (DCDS-B)*, 1(1):89–102, 2001.

[41] D. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order*. Classics in Mathematics. Springer-Verlag, Berlin, Heidelberg, 2nd edition, 2001.

[42] H. Glauert. The analysis of experimental results in the windmill brake and vortex ring states of an airscrew. *London: Aeronautical Research Committee*, 1026, 1926.

[43] H. Glauert. Airplane propellers. In W. F. Durand, editor, *Aerodynamic Theory*, volume 4, pages 169–360. Berlin: Julius Springer, 1935.

[44] H. Glauert. *The Elements of Aerofoil and Airscrew Theory*. Cambridge Science Classics. Cambridge University Press, 1983.

[45] I. G. Graham, O. R. Pembery, and E. A. Spence. The Helmholtz equation in heterogeneous media: a priori bounds, well-posedness, and resonances. *ArXiv e-prints*, 2018.

[46] I. G. Graham and S. A. Sauter. Stability and error analysis for the Helmholtz equation with variable coefficients. *ArXiv e-prints*, 2018.

[47] M. O. Hansen. *Aerodynamics of Wind Turbines*. Taylor and Francis, 2015.

[48] J. Haslinger and R. A. E. Mäkinen. On a topology optimization problem governed by two-dimensional Helmholtz equation. *Computational Optimization and Applications*, 62(2):517–544, 2015.

[49] U. Hetmaniuk. Stability estimates for a class of Helmholtz problems. *Communications in Mathematical Sciences*, 5(3):665–678, 2007.

[50] B. D. Hibbs. HAWT performance with dynamic stall. Technical report, Solar Energy Research Institute, 1986.

[51] M. Honnorat, J. Monnier, and F.-X. Le Dimet. Lagrangian data assimilation for river hydraulics simulations. *Computing and Visualization in Science*, 12(5):235–246, 2009.

[52] X. Hu, F. Sun, and Y. Zou. Estimation of state of charge of a lithium-ion battery pack for electric vehicles using an adaptive Luenberger observer. *Energies*, 3:1586–1603, 2010.

[53] K. Ide, P. Courtier, M. Ghil, and A. C. Lorenc. Unified notation for data assimilation : Operational, sequential and variational. *Journal of the Meteorological Society of Japan. Ser. II*, 75(1B):181–189, 1997.

[54] G. Ingram. Wind turbine blade analysis using the blade element momentum method, version 1.1., 2011.

[55] D. Isebe, P. Azerad, B. Mohammadi, and F. Bouchette. Optimal shape design of defense structures for minimizing short wave impact. *Coastal Engineering*, 55(1):35–46, 2008.

[56] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transaction of the ASME—Journal of Basic Engineering*, 82(1):35–45, 1960.

[57] O. A. Ladyzhenskaya and N. N. Ural'tseva. *Linear and quasilinear elliptic equations*, volume 46 of *Mathematics in Science and Engineering*. Academic Press, New York, 1968.

[58] F.-X. Le Dimet and O. Talagrand. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A*, 38A(2):97–110, 1986.

[59] B. Le Méhauté. *An Introduction to Hydrodynamics and Water Waves*. Springer Study Edition. Springer-Verlag, New York, 1976.

[60] J.-L. Lions. *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*. Etudes Mathématiques. Dunod, 1968.

[61] J.-L. Lions. On the Schwarz alternating method. I. In R. Glowinski, G. H. Golub, G. A. Meurant, and J. Périaux, editors, *First International Symposium on Domain Decomposition Methods for Partial Differential Equations*, pages 1–42. SIAM, 1988.

[62] J.-L. Lions. On the Schwarz alternating method II: Stochastic interpretation and orders properties. In J. P. Tony Chan, Roland Glowinski and O. Widlund, editors, *Domain Decomposition Methods*, page 47–70. SIAM, 1989.

[63] J.-L. Lions. On the Schwarz alternating method. III: a variant for nonoverlapping subdomains. In J. P. Tony Chan, Roland Glowinski and O. Widlund, editors, *Third International Symposium on Domain Decomposition Methods for Partial Differential Equations*, pages 202–223. SIAM, 1989.

[64] J.-L. Lions, Y. Maday, and G. Turinici. Résolution d'EDP par un schéma en temps «pararéel». *Comptes Rendus de l'Académie des Sciences - Série I - Mathématique*, 332(7):661–668, 2001.

[65] D. Luenberger. *Introduction to Dynamic Systems: Theory, Models, and Applications*. John Wiley & Sons, New York, 1979.

[66] Y. Maday, J. Salomon, and G. Turinici. Monotonic parareal control for quantum systems. *SIAM Journal on Numerical Analysis*, 45(6):2468–2482, 2007.

[67] D. Maniaci. *49th AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition*, chapter An Investigation of WT_Perf Convergence Issues. Aerospace Sciences Meetings. American Institute of Aeronautics and Astronautics, Jan 2011. 0.

[68] J. F. Manwell, J. G. McGowan, and A. L. Rogers. *Wind Energy Explained: Theory, Design and Application, Second Edition*, volume 30. John Wiley and Sons, Ltd, 03 2006.

[69] B. Mohammadi and A. Bouharguane. Optimal dynamics of soft shapes in shallow waters. *Computers and Fluids*, 40(1):291–298, 2011.

[70] C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49, 2003.

[71] H. Nersisyan, D. Dutykh, and E. Zuazua. Generation of two-dimensional water waves by moving bottom disturbances. *IMA Journal of Applied Mathematics*, 80(4):1235–1253, 2014.

[72] J. Nievergelt. Parallel methods for integrating ordinary differential equations. *Commun. ACM*, 7(12):731–733, 1964.

[73] R. Nittka. Regularity of solutions of linear second order elliptic and parabolic boundary value problems on Lipschitz domains. *Journal of Differential Equations*, 251:860–880, 2011.

[74] M. Nodet. *Inverse problems for the environment: tools, methods and applications*. PhD thesis, Université de Grenoble, 2013.

[75] J.-C. Nédélec. *Acoustic and Electromagnetic Equations: Integral Representations for Harmonic Problems*, volume 144 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2001.

[76] G. Pagès, O. Pironneau, and G. Sall. The parareal algorithm for american options. *Comptes Rendus Mathematique*, 354(11):1132 – 1138, 2016],.

[77] L. Prandtl and A. Betz. Vier abhandlungen zur hydrodynamik und aerodynamik. *Göttinger Nachr.*, pages 88–92, 1927.

[78] W. J. M. Rankine. On the mechanical principles of the action of propellers. *Transactions, Institute of Naval Architects*, 6:13–30, 1865.

[79] V. Rao and A. Sandu. A time-parallel approach to strong-constraint four-dimensional variational data assimilation. *Journal of Computational Physics*, 313:583–593, 2016.

[80] J. Sainte-Marie. Vertically averaged models for the free surface Euler system. derivation and kinetic interpretation. *Mathematical Models and Methods in Applied Sciences (M3AS)*, 21(3):459–490, 2011.

[81] Y. Sasaki. An objective analysis based on the variational method. *Meteorological Society of Japan*, 36(3):77–88, 1958.

[82] H. A. Schwarz. Über einen grenzübergang durch alternierendes verfahren. *Vierteljahrsschrift der Naturforschenden Gesellschaft in Zurich*, 15:272–286, 1870.

[83] M. Sellier. Inverse problems in free surface flows: a review. *Acta Mechanica*, 227(3):913–935, 2016.

[84] W. Z. Shen, R. Mikkelsen, J. N. Sorensen, and C. Bak. Tip loss corrections for wind research turbine computations. *WIND ENERGY*, 8(4):457–475, OCT-DEC 2005.

[85] Q. Song and W. D. Lubitz. BEM simulation and performance analysis of a small wind turbine rotor. *Wind Engineering*, 37(4):381–399, 2013.

[86] J. Sørensen. *General Momentum Theory for Horizontal Axis Wind Turbines*. Springer, 2016.

[87] D. Spera, editor. *Wind Turbine Technology: Fundamental Concepts in Wind Turbine Engineering, Second Edition*. ASME, New York, NY, 2009.

[88] Y. Trémolet and F.-X. Le Dimet. Parallel algorithms for variational data assimilation and coupling models. *Parallel Computing*, 22(5):657–674, 1996.

[89] F. Ursell. The long-wave paradox in the theory of gravity waves. *Mathematical Proceedings of the Cambridge Philosophical Society*, 49(4):685–694, 1953.

[90] C. Werndl. Initial-condition dependence and initial-condition uncertainty in climate science. *The British Journal for the Philosophy of Science*, 2018.

[91] R. E. Wilson, P. B. S. Lissaman, and S. N. Walker. Aerodynamic performance of wind turbines. final report. Technical report, Oregon State University, 1976.

RÉSUMÉ

La présente thèse vise à contribuer à l'élaboration d'un cadre théorique pour trois problèmes dans le contexte des énergies marines renouvelables. Dans sa première partie, nous proposons une procédure pour coupler des méthodes d'assimilation de données temporelles non limitées avec des algorithmes parallèles en temps. La combinaison entre l'observateur de Luenberger et l'algorithme Parareél est étudiée, ce qui permet d'estimer le nombre d'itérations parareéelles nécessaires pour préserver le taux de convergence de l'observateur et d'obtenir une estimation de l'efficacité théorique de l'ensemble de la procédure.

Nous discutons ensuite la détermination d'une bathymétrie dans une perspective d'optimisation. En imposant que la propagation des vagues optimise un certain critère associé à une fonctionnelle de coût, nous considérons un problème d'optimisation sous contrainte d'EDP où la bathymétrie joue le rôle de contrôle et la propagation des vagues est décrite par une équation de type Helmholtz. Nous sommes en mesure de prouver, sur la base d'hypothèses appropriées, la continuité de la fonction contrôle-état et l'existence d'une solution optimale, incluant aussi quelques résultats sur les solutions au problème de Helmholtz et la convergence dans un cadre discret. Ce travail est complété par des expériences numériques.

La dernière partie de ce travail est consacrée à l'analyse de la méthode de l'élément de pale (BEM), une méthode classique utilisée pour déterminer les performances d'une hélice ainsi que des paramètres de design. Nous proposons une reformulation de la méthode qui permet d'obtenir des conditions d'existence des solutions et d'établir la convergence de certains algorithmes de résolution. Nous étudions également le problème d'optimisation associé dans certains contextes.

MOTS CLÉS

Algorithme parareél, Assimilation parallèle de données, Optimisation d'une bathymétrie, Équation de Helmholtz, Méthode de l'élément de pale, Design des pales.

ABSTRACT

The present thesis aims to contribute to the development of a theoretical framework for three problems in the context of renewable marine energy. In the first part, we propose a procedure to couple unbounded in time data assimilation methods with time-parallel algorithms. The combination between the Luenberger observer and Parareal algorithm is studied, providing a way to estimate the number of parareal iterations required to preserve the observer rate of convergence, as well as an estimation of the theoretical efficiency of the entire procedure.

We then discuss the determination of a bathymetry from an optimization perspective. Imposing that wave propagation must fulfill a certain criterion associated with a cost functional, we consider a PDE-constrained optimization problem where the bathymetry plays the role of control and wave propagation is described by the Helmholtz equation. We are able to prove, under suitable assumptions, the continuity of the control-to-state mapping and the existence of an optimal solution, including also some results about solutions to Helmholtz problem and convergence in a discrete framework. This work is complemented by numerical experiments.

The last part of this work is devoted to analyze the convergence of the Blade element momentum (BEM) theory, a classical method used to determine the propeller efficiency as well as its design parameters. We propose a reformulation of the method that allows to obtain conditions for existence of solutions and establish the convergence of some solving algorithms. We also study the associated optimization problem in certain contexts.

KEYWORDS

Parareal algorithm, Parallel data assimilation, Bathymetry optimization, Helmholtz equation, Blade element method, Blade design.